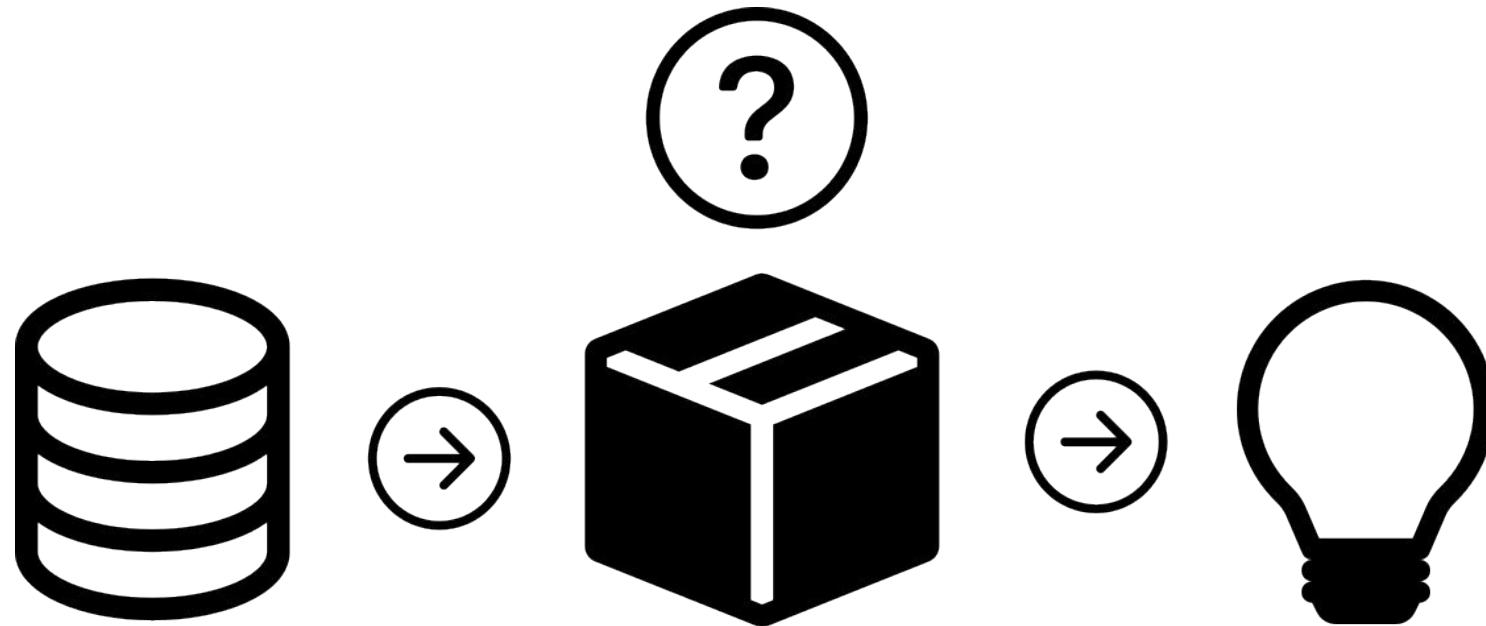


Unveiling the black box: Explainability on Machine Learning Techniques

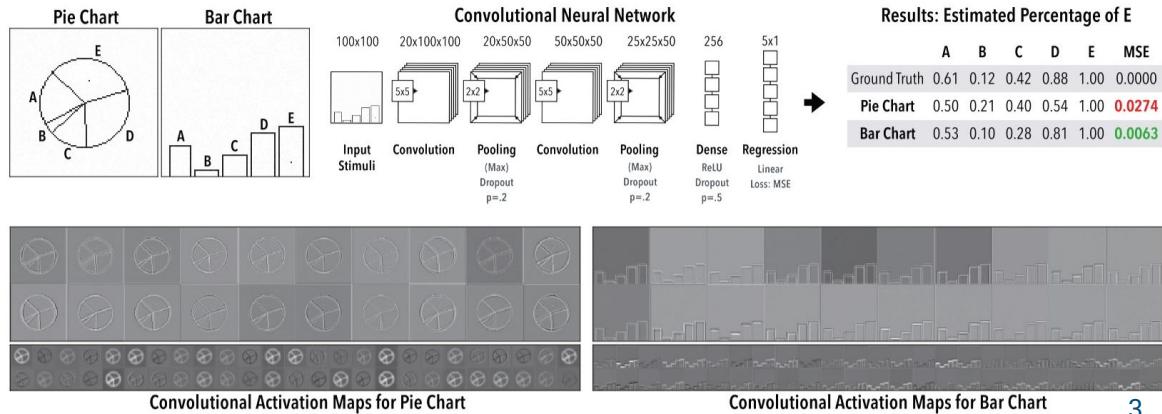
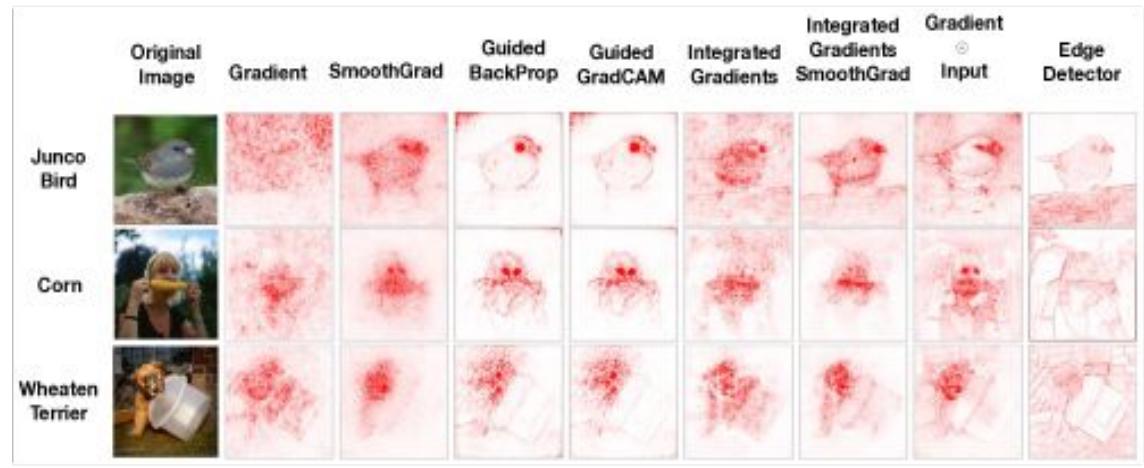
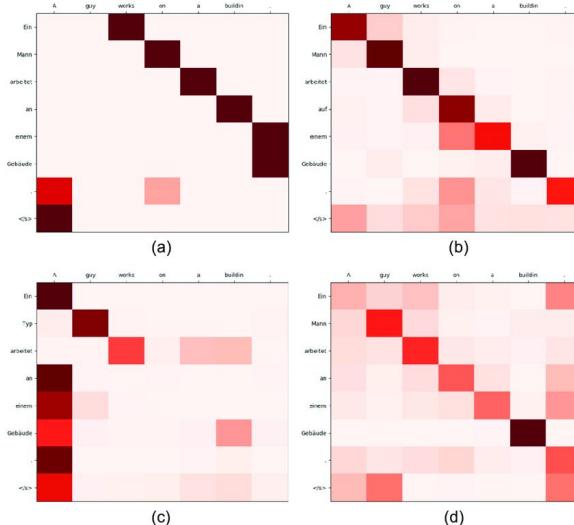
Tiago Araújo
Researcher

Introduction

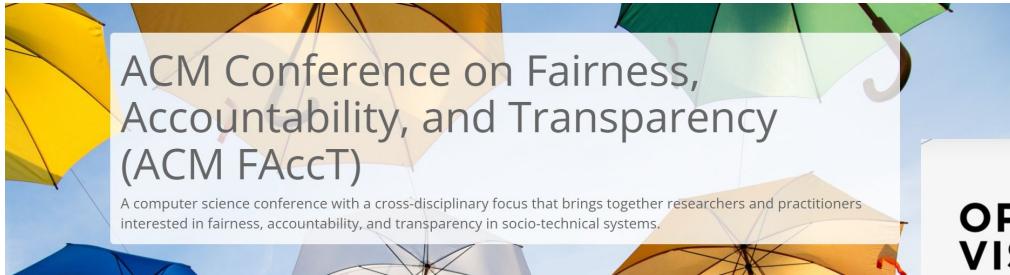


Good Motivators

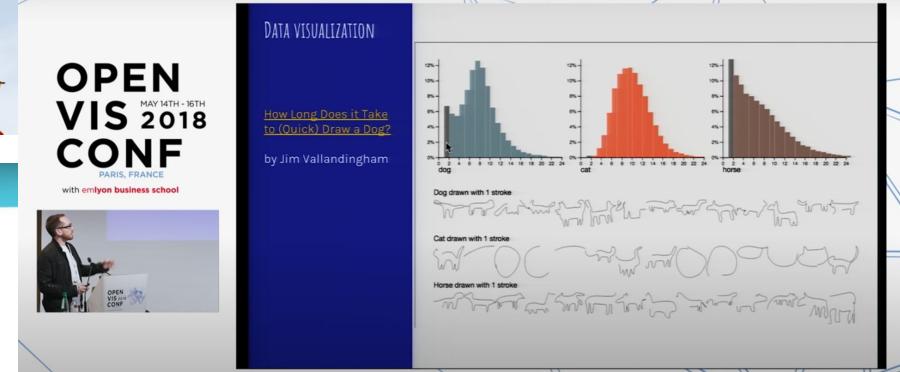
—



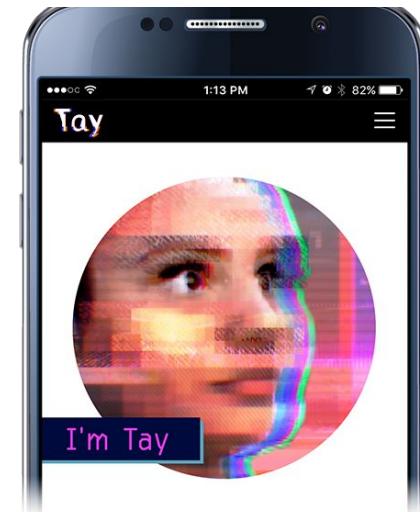
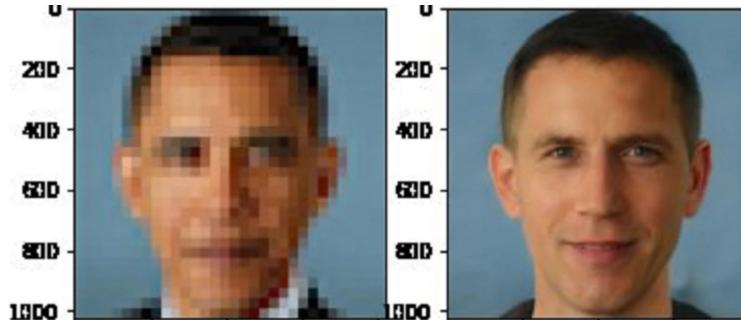
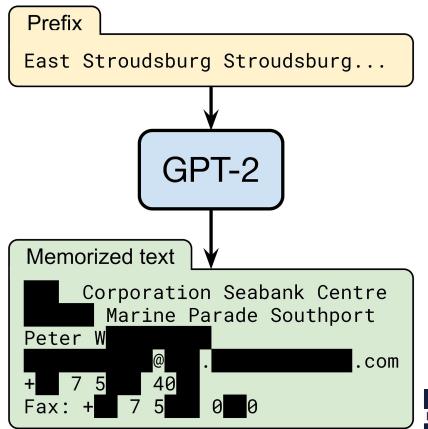
Good Motivators



A composite image. On the left, a blue banner for the OPEN VIS 2018 CONF in Paris, France, featuring a red geometric logo and the text "OPEN VIS 2018 CONF PARIS, FRANCE with emlyon business school". In the center, two AI-generated images are shown: a "Frog" image labeled "100% confident" and a "Bird" image labeled "99.9% confident", both consisting of a dense, multi-colored pixelated pattern. On the right, a slide titled "1. Rubbish Examples: optimized for confidence" shows a speaker at a podium and a histogram titled "How Long Does it Take to (Quick) Draw a Dog?" by Jim Vallandingham. The histogram shows the distribution of stroke times for drawing dogs, cats, and horses.



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



FEDERAL TRADE COMMISSION
PROTECTING AMERICA'S CONSUMERS

Enforcement ▾ Policy ▾ Advice and Guidance ▾ News and Events ▾ About the FTC ▾ Q

[Home](#) / [Legal Library](#) / [Browse](#) / [Statutes](#)

Federal Trade Commission Act



Tags: Consumer Protection | Competition | Appliances | Alcohol | Automobiles | Clothing and Textiles | Finance | Franchises, Business Opportunities, and Investments | Jewelry | Real Estate and Mortgages | Tobacco | Advertising and Marketing | Children's Products, Endorsements, Influencers, and Reviews | Environmental Marketing | Health Claims | Made in USA | Online Advertising and Marketing | Telemarketing | Advertising and Marketing Basics | Credit and Finance | Credit and Loans | Debt | Debt Collection | Mortgages | Payments and Billing | Privacy and Security | Children's Privacy | Consumer Privacy | Credit Reporting | Data Security | Gramm-Leach-Bliley Act | Red Flags Rule

ACM TechTalks

ACM TechTalks on Artificial Intelligence & Machine Learning

Beyond the Data

Data protection is not enough

[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](#)

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

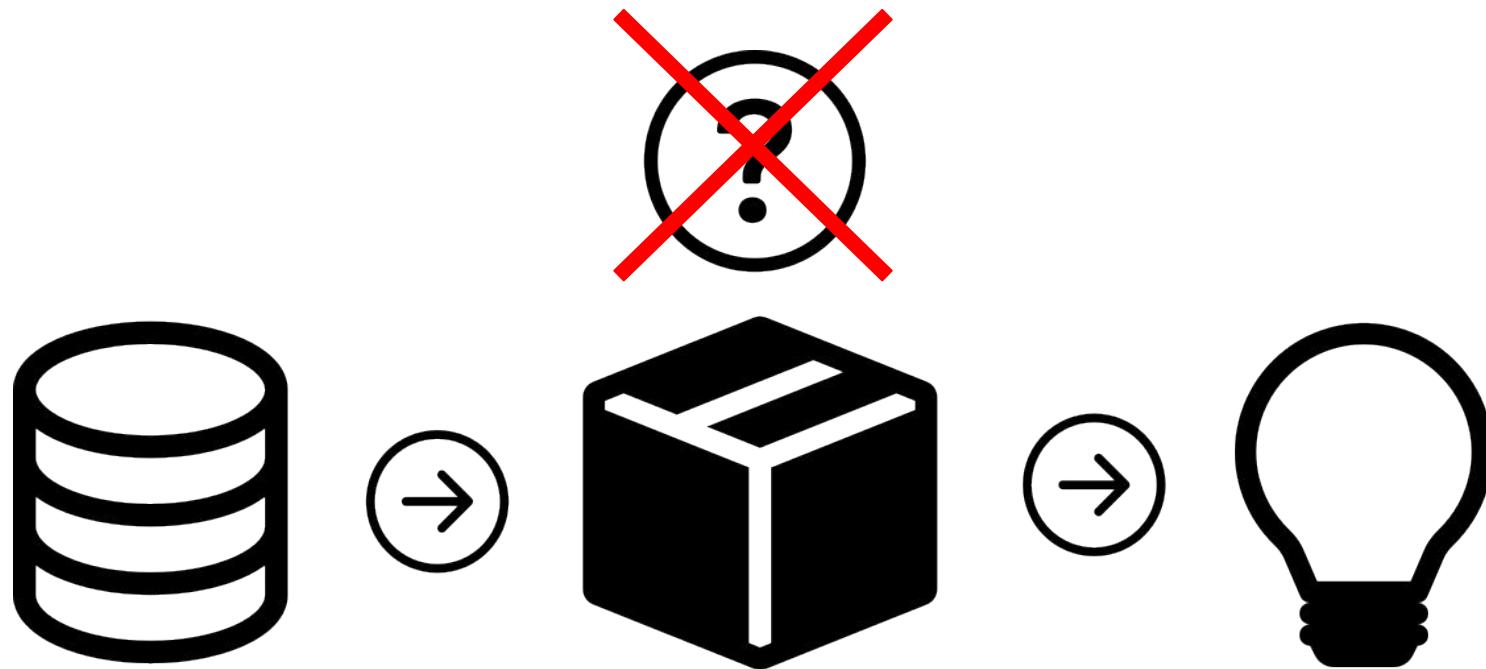
IN THE UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF DELAWARE

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

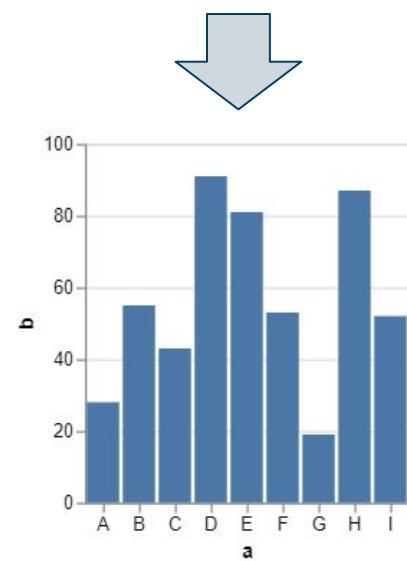
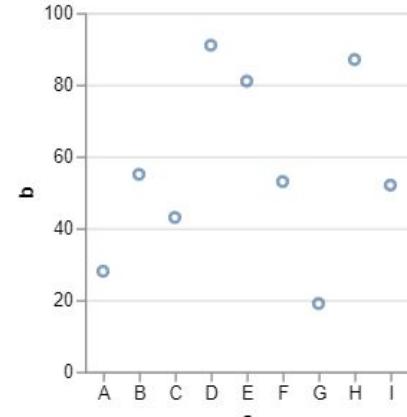
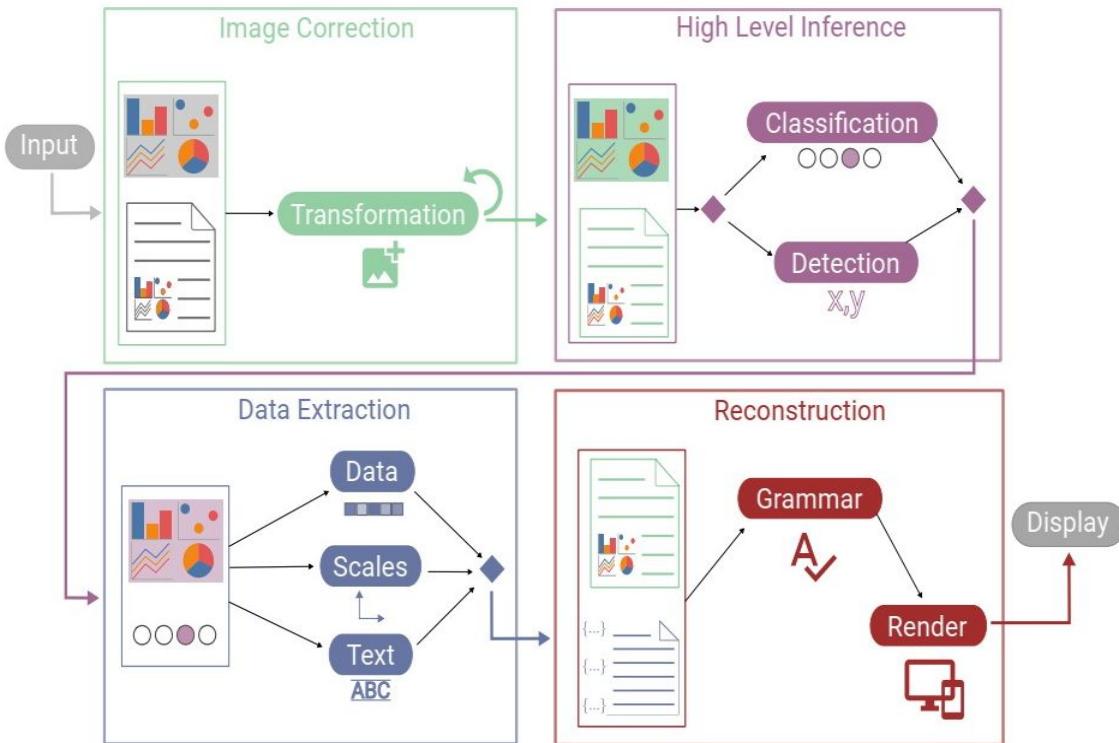
COMPLAINT

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:

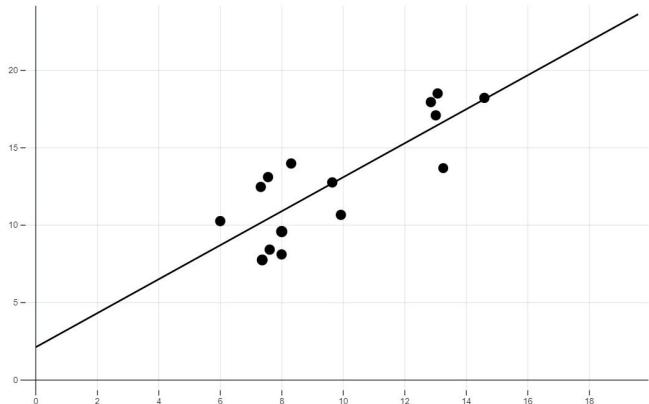
Explainable and Interpretable AI



Me and ML



History

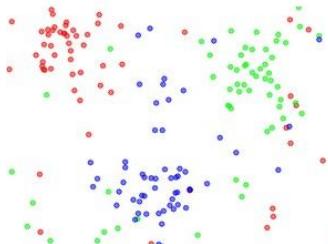


		SPRINKLER	
RAIN	T	F	
F	0.4	0.6	
T	0.01	0.99	

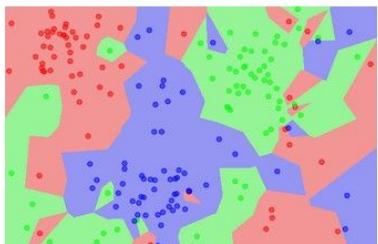


SPRINKLER	RAIN	GRASS WET	
		T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

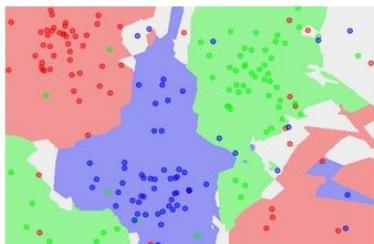
the data



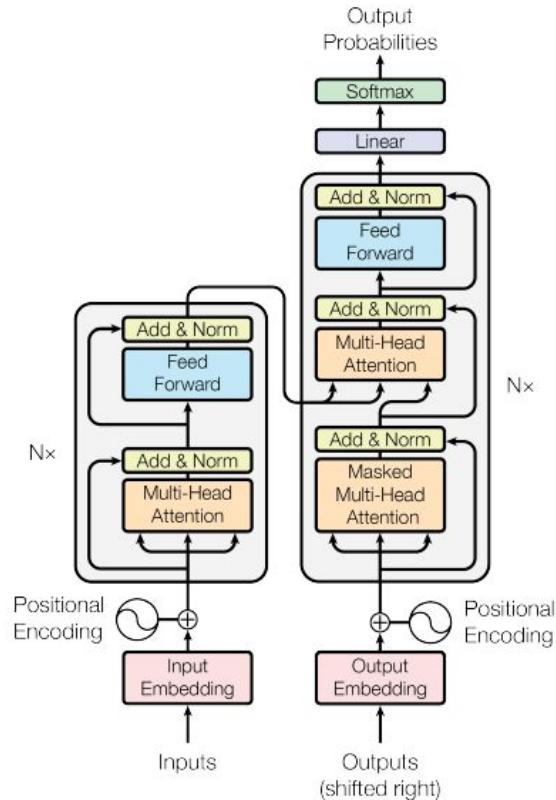
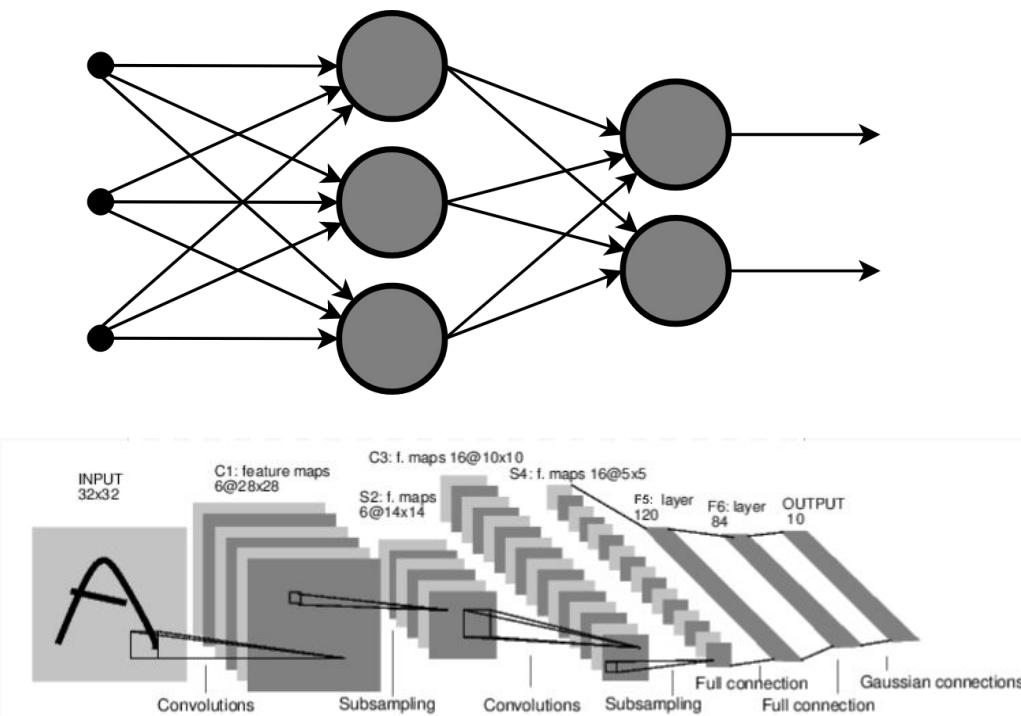
NN classifier



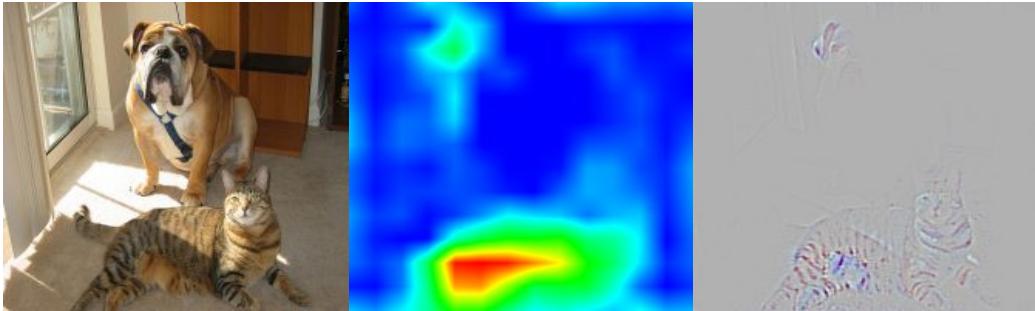
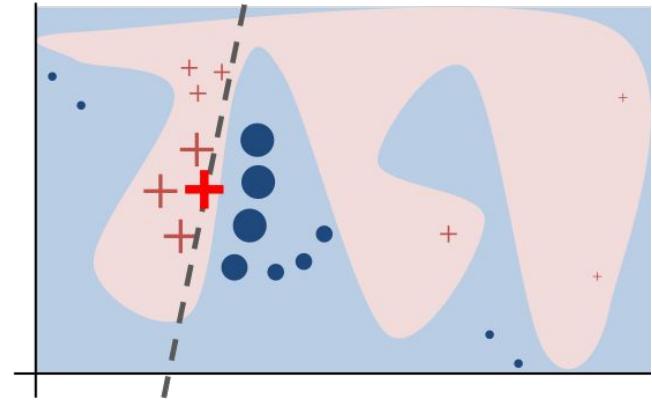
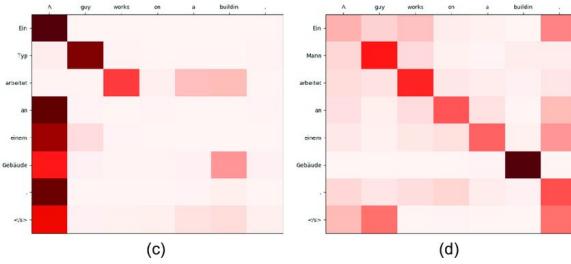
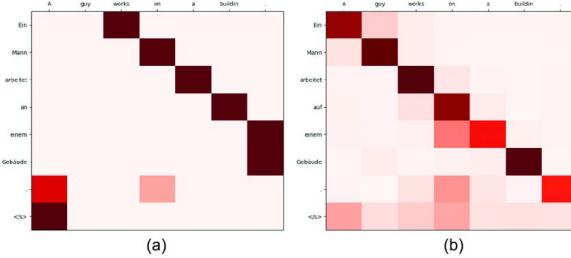
5-NN classifier



ML Models



History

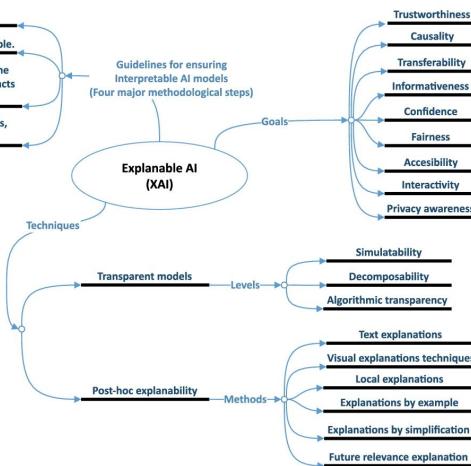


Abstractions

Contextual factors, potential impacts and domain-specific needs must be taken into account when devising an approach to interpretability.

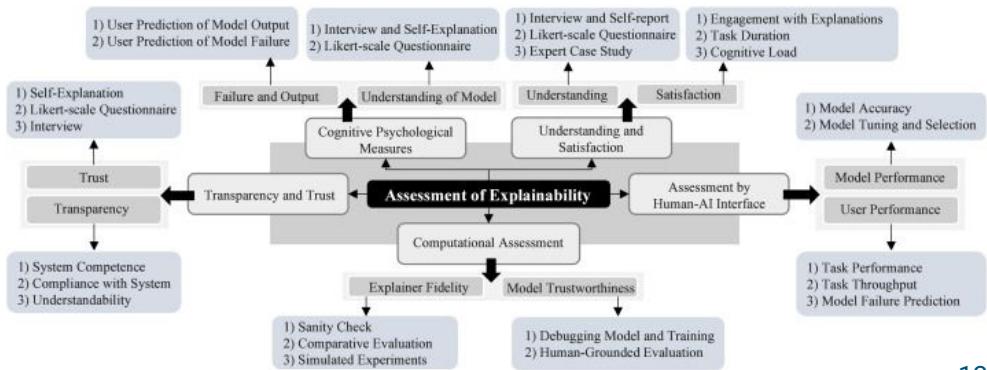
Interpretable techniques should be preferred when possible.

To rethink interpretability in terms of the cognitive skills, competencies and limitations of the individual human



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

[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



Accountability

Privacy

Transparency

Environmental Impact & Sustainability

Accuracy

Fairness

Human Control and Decision-making

Bias

Safety & Security

[nature](#) > [humanities and social sciences communications](#) > [articles](#) > [article](#)

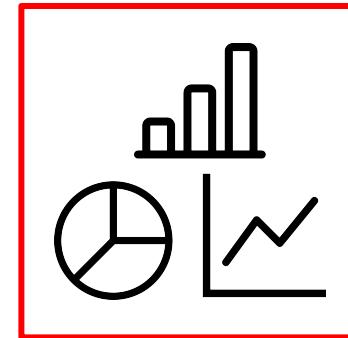
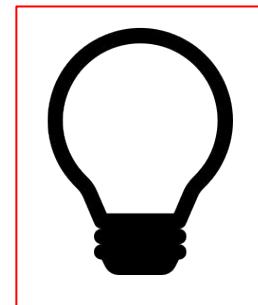
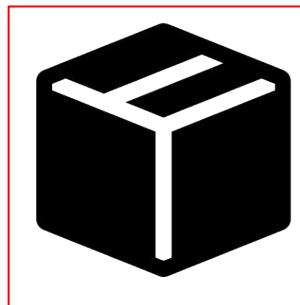
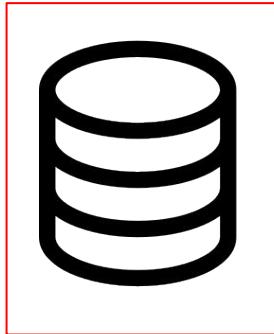
Article | [Open Access](#) | Published: 17 June 2020

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

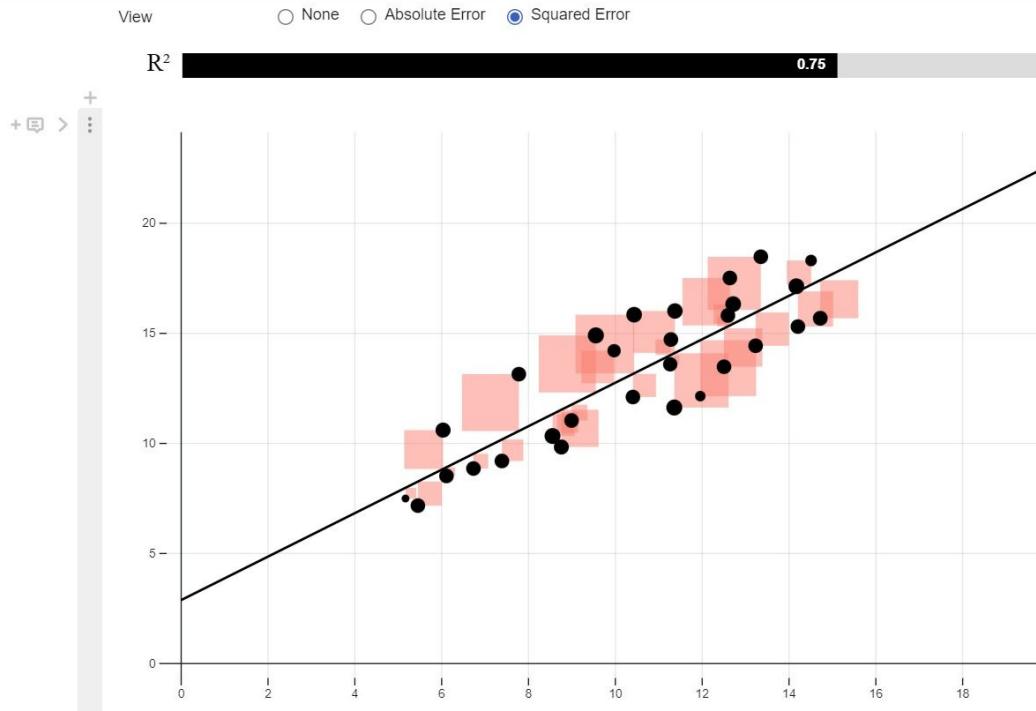
Samuele Lo Piano

Methods

$\Sigma \pi$

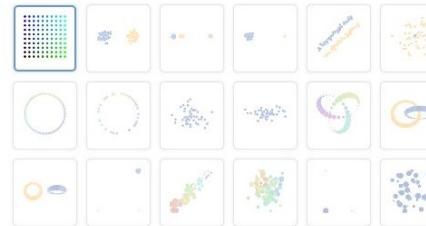
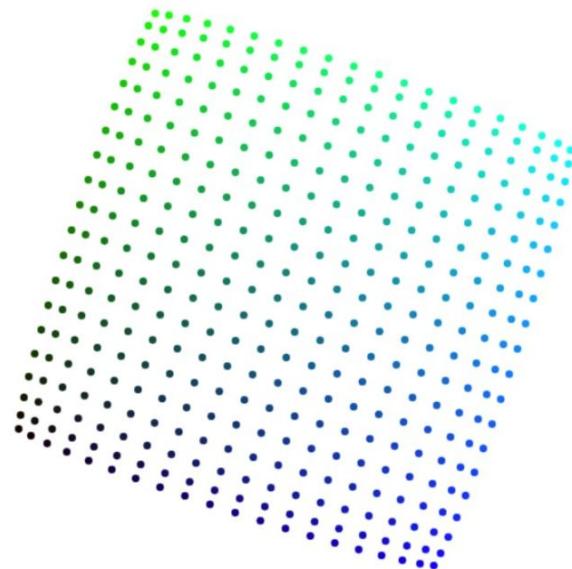


Linear Regression



[Interactive Visualization of Linear Regression](#)

Data



Step
2,070

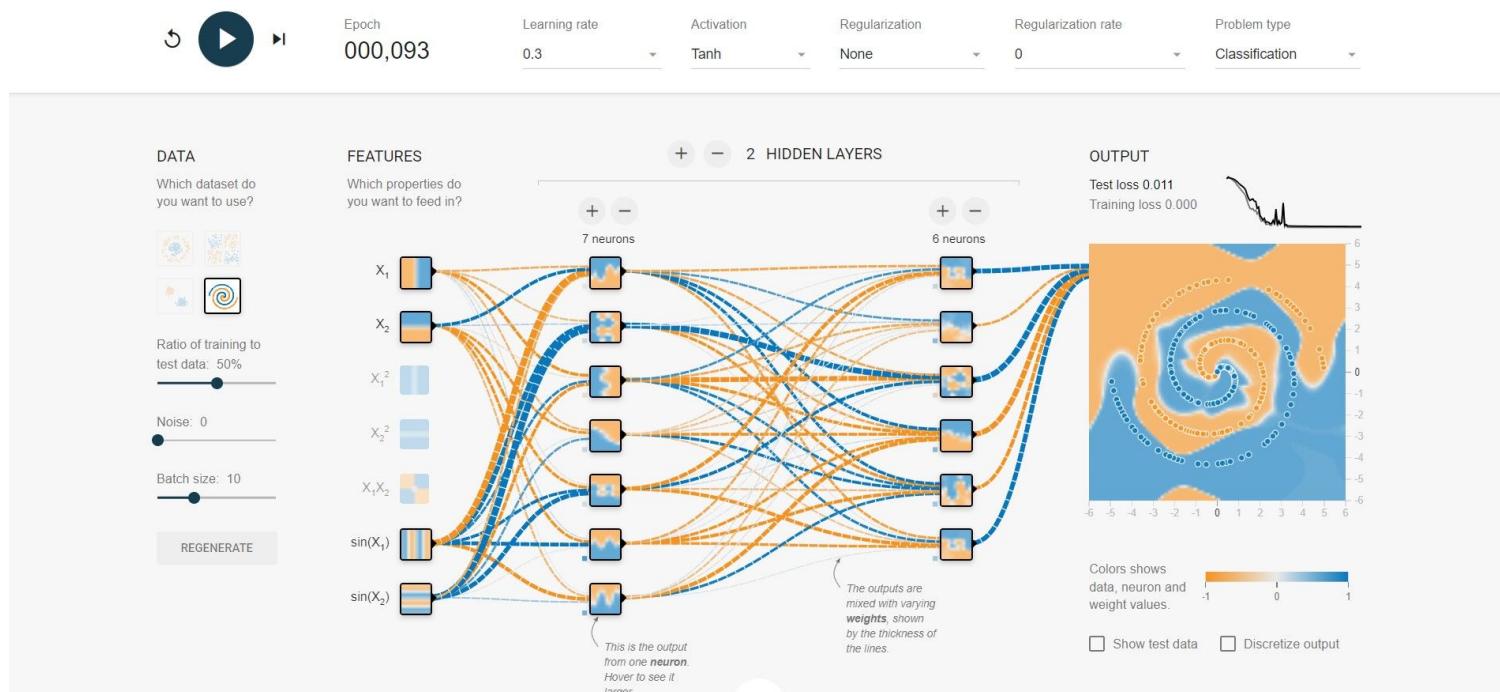
Points Per Side 20

Perplexity 10

Epsilon 5

A square grid with equal spacing between points.
Try convergence at different sizes.

Neural Networks



A Neural Network Playground

Clustering

EduClust

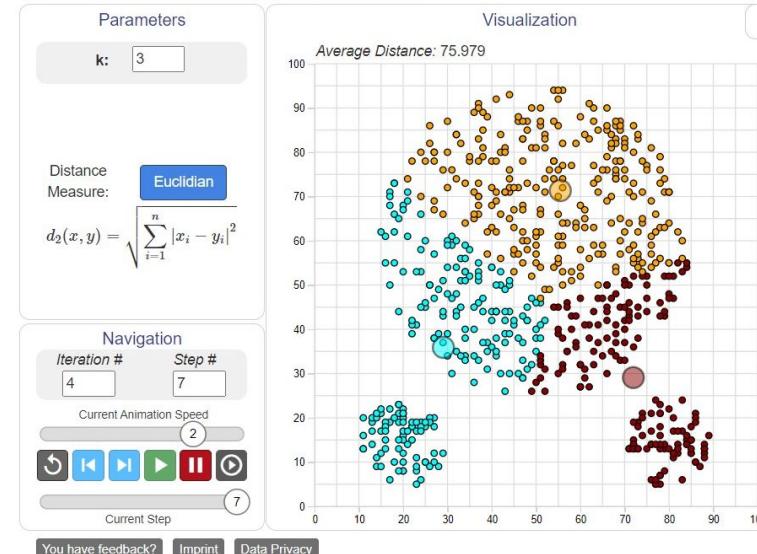
A Visual Education Platform
for Teaching Clustering Algorithms

Johannes Fuchs, Petra Isenberg, Anastasia Bezerianos, Matthias Müller, Daniel Keim

k-means

Three Not Equal Circles

Custom Data



Data Analysis and Visualization



Pseudocode: k-means

Complexity Range: O($k \times n \times t$)

Input: k clusters

Output: k clusters

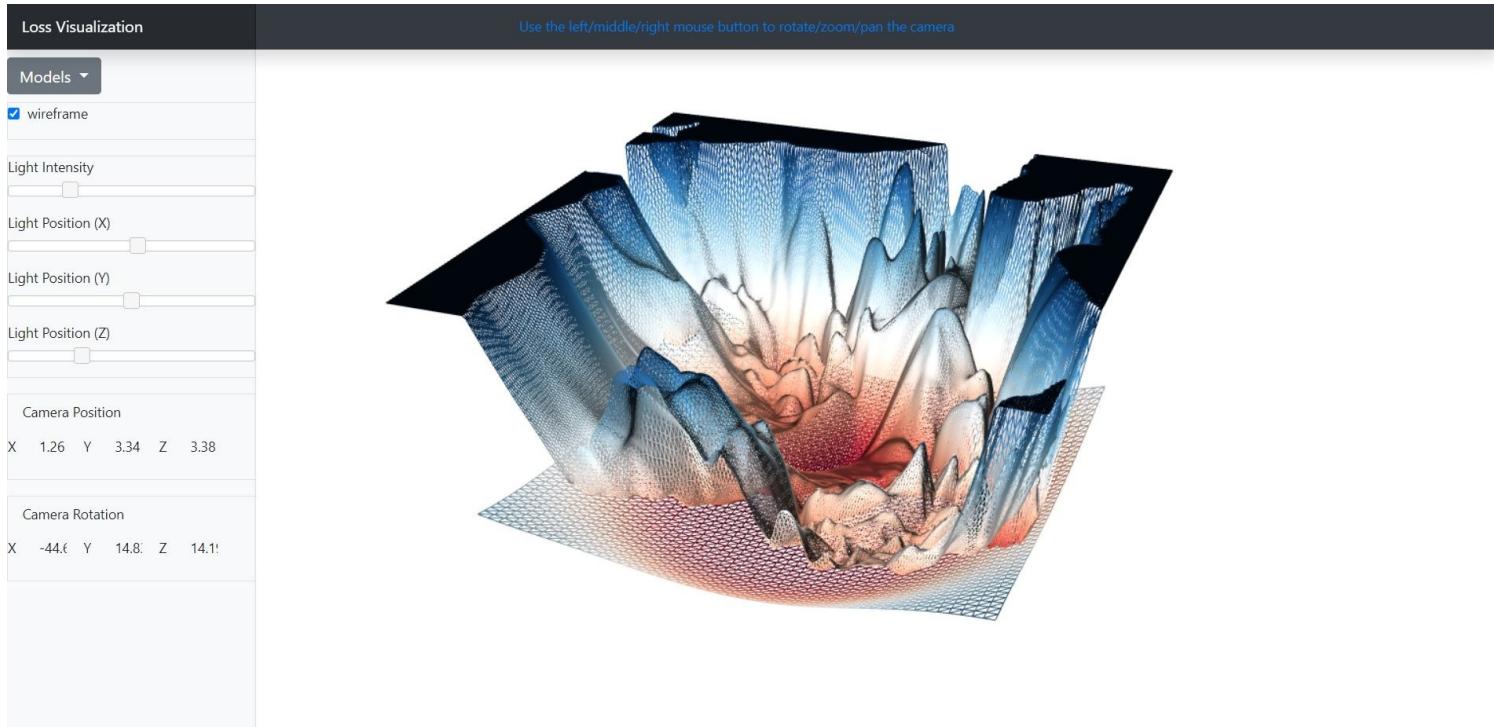
Pseudocode:

1. Choose k objects as initial **cluster centers**.
2. Assign each data point to the cluster which has the closest **mean point (centroid)** under chosen distance metric.
3. When all data points have been assigned, recalculate the positions of k **centroids (mean points)**.
4. Repeat steps 2 and 3 until the **centroids** do not change any more. All data points remain in their most recently assigned cluster.

Universität
Konstanz



Loss Function



Visualizing the Loss Landscape of Neural Nets

Convolutional Neural Networks

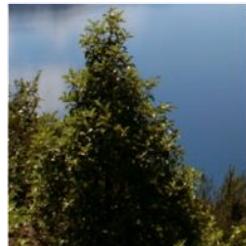


[ConvNetJS CIFAR-10 demo](#)

Convolutional Neural Networks



Horizon



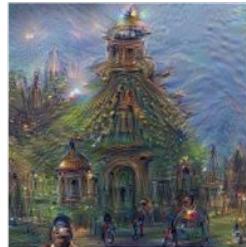
Trees



Leaves



Towers & Pagodas



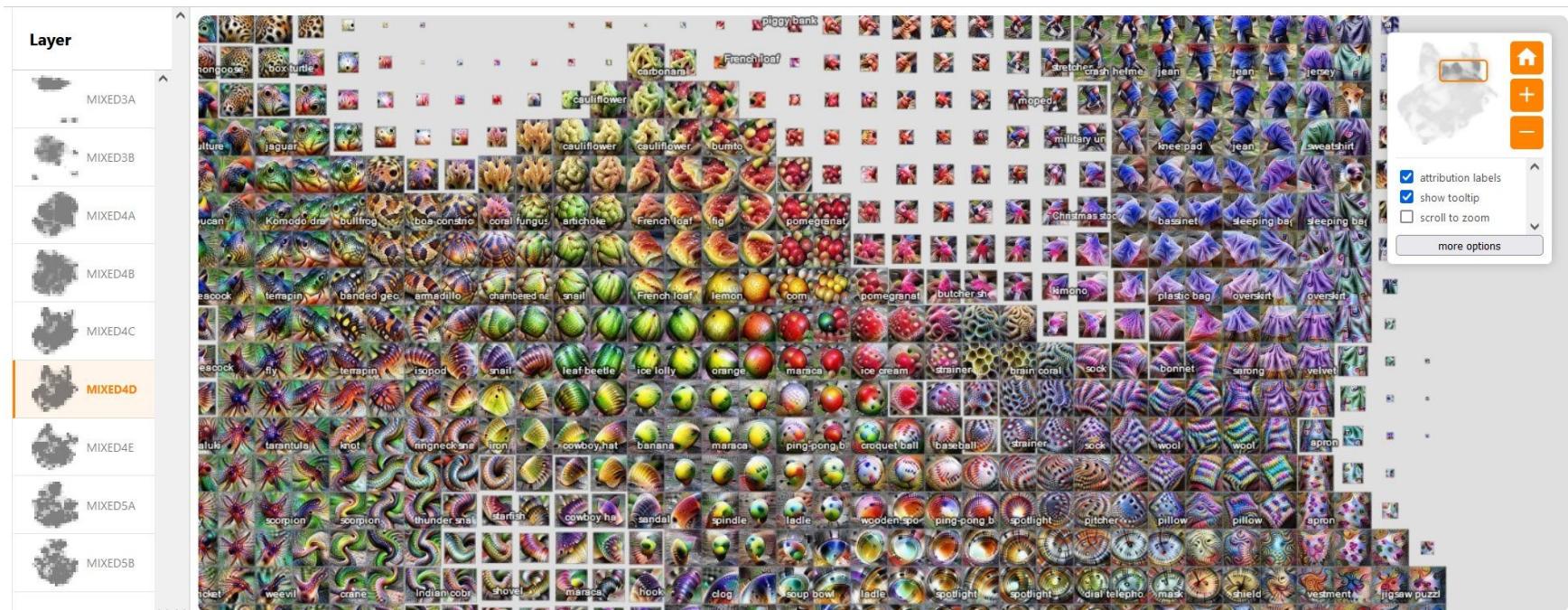
Buildings



Birds & Insects

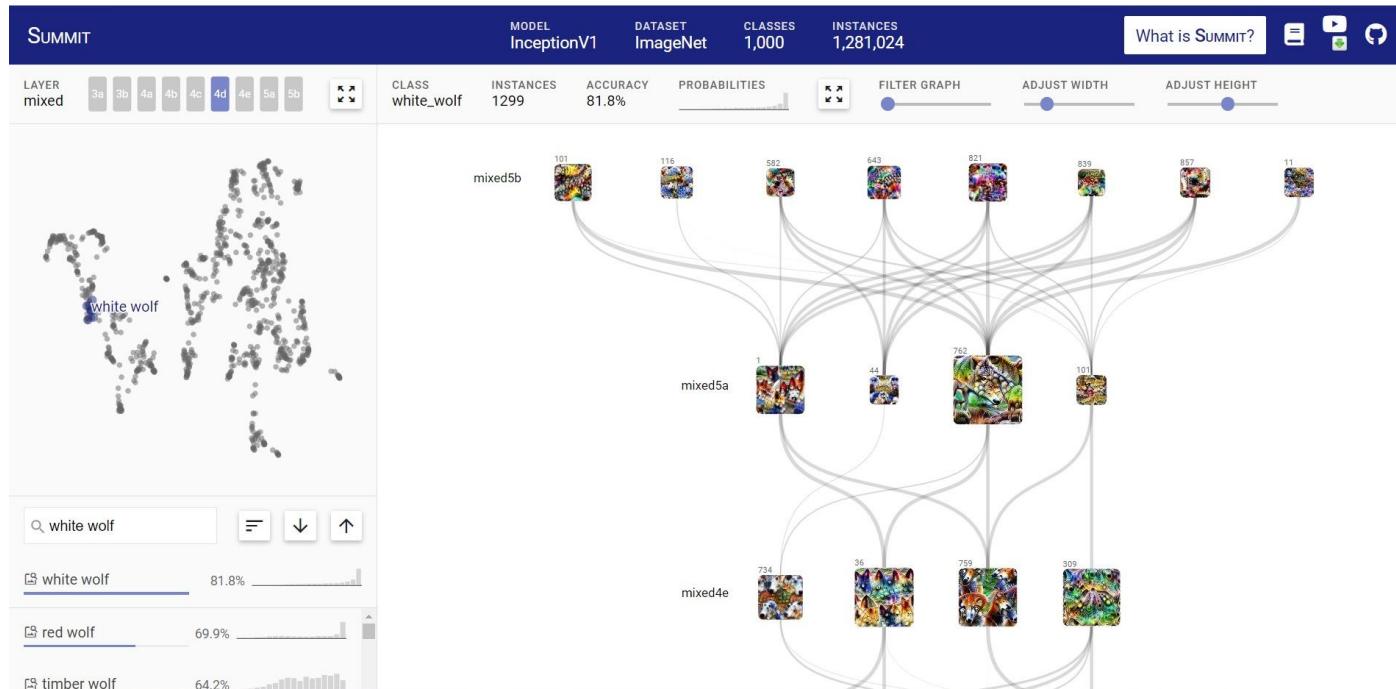
Inceptionism: Going Deeper into Neural Networks

Activation Maps



Exploring Neural Networks with Activation Atlases

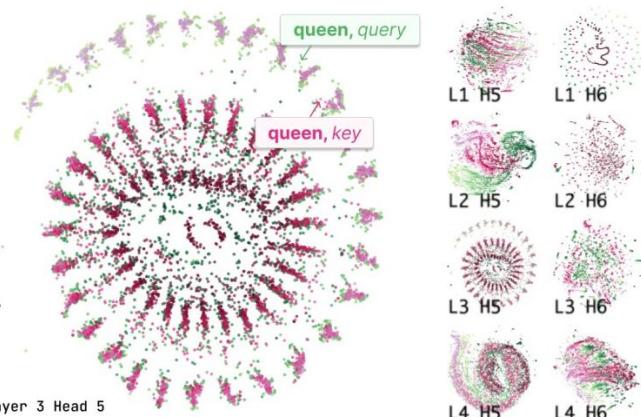
Activation Maps



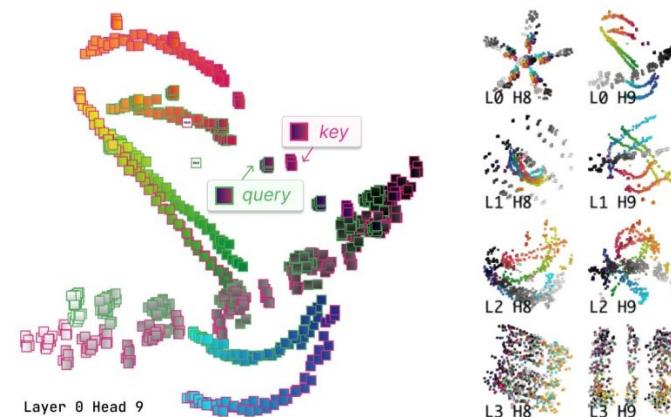
[Summit: Scaling Deep Learning Interpretability by Visualizing Activation and Attribution Summarizations](#)

Transformers

(a) Language Transformer



(b) Vision Transformer



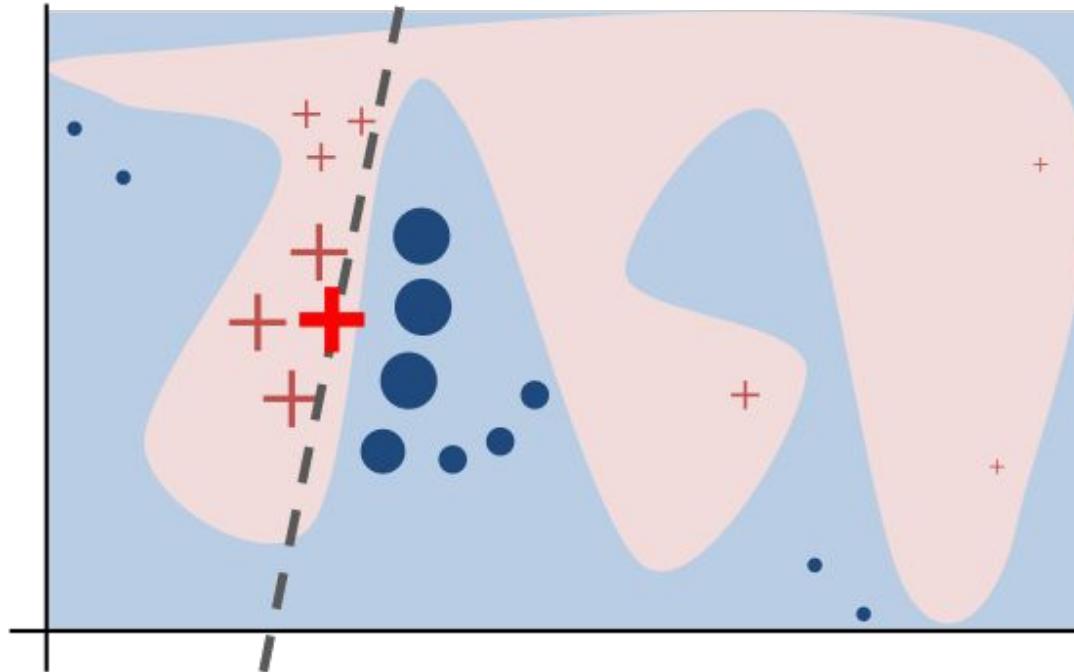
(c) Source Sentences (Sample):

"Plumb was awarded a knighthood in the Queen's Birthday Honours list in 1973."
"This is estimated to make up between 5% and 72% of cases."
"He read and memorized the entire Quran by the time he was nine years old."

(d) Source Images:

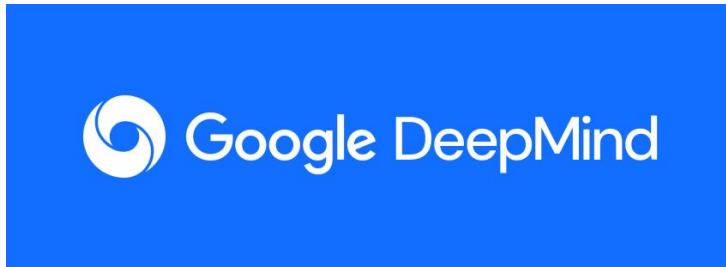


Input and predictions

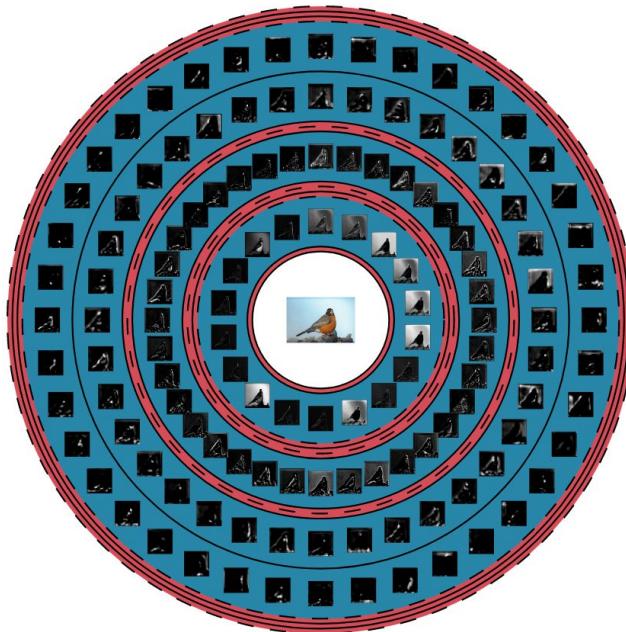
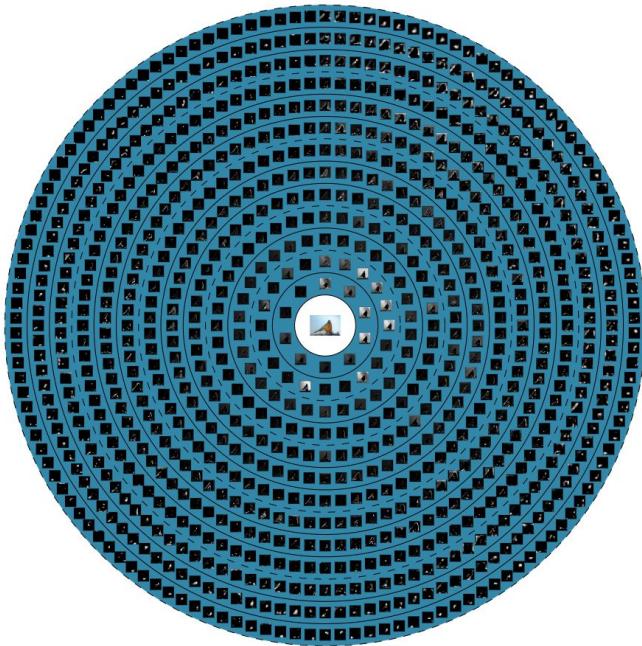


Lime: Explaining the predictions of any machine learning classifier

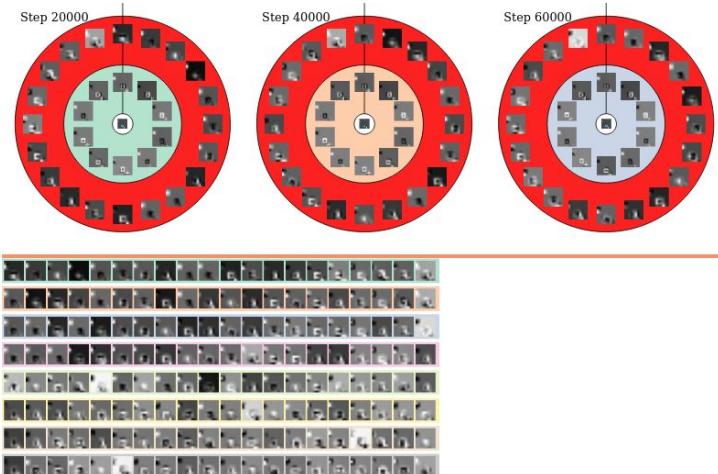
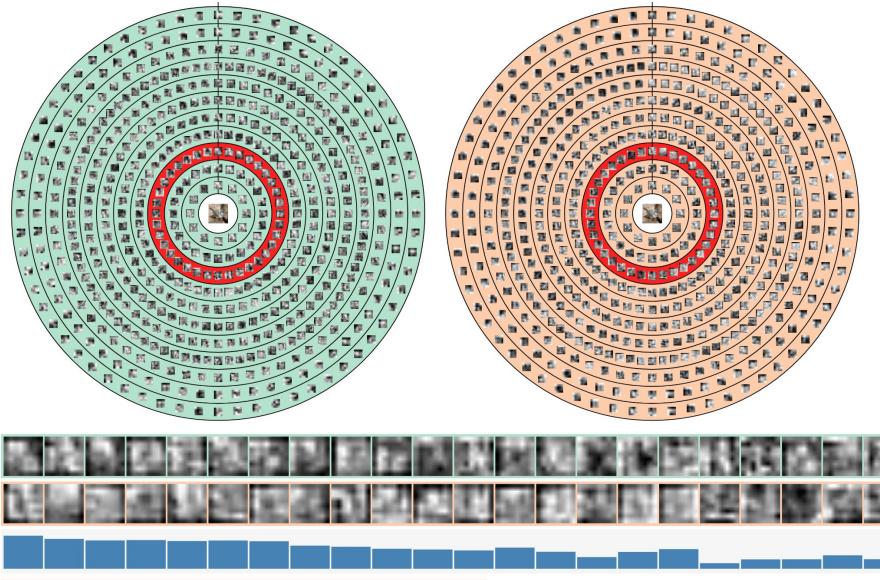
Institutions



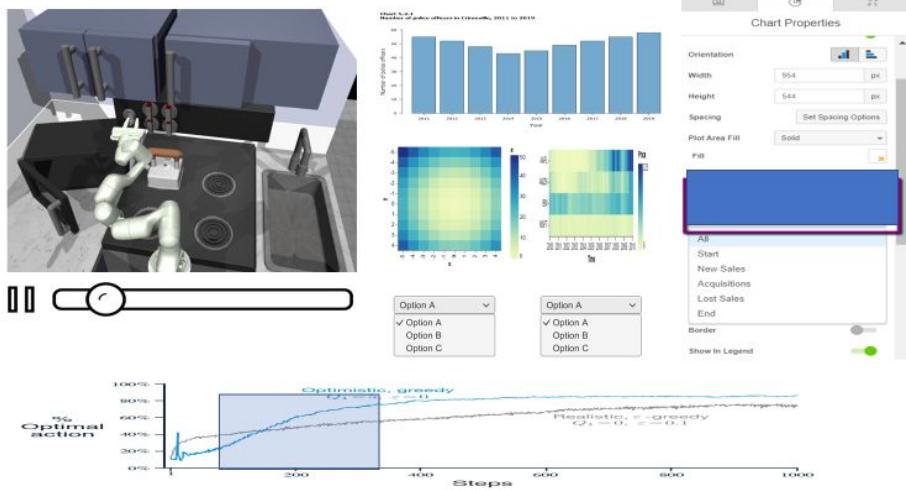
Our contribution



Our contribution



Our future contribution





Challenges

Human in the loop

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 ▾



An armchair in the shape of an avocado



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>

