Interpretability of ML Systems

Philipp @ DSR

Class Outline

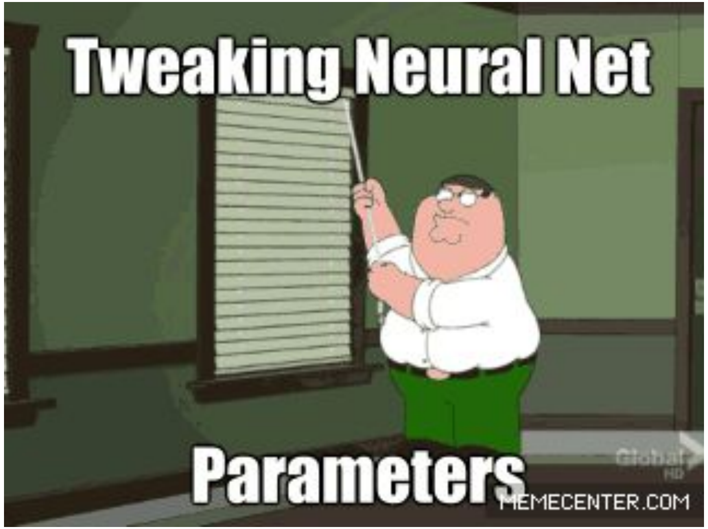
- Day 1
  - Introduction to Interpretability
  - Paper session
  - **Lunch**
  - Tooling for interpretability (Notebooks/Exercises)
  - Case study for vision/text based models

Class Outline

- Day 1
  - Introduction to Interpretability
  - Paper session
  - **Lunch**
  - Tooling for interpretability (Notebooks/Exercises)
  - Case study for vision/text based models
- Day 2
  - Support for capstone project

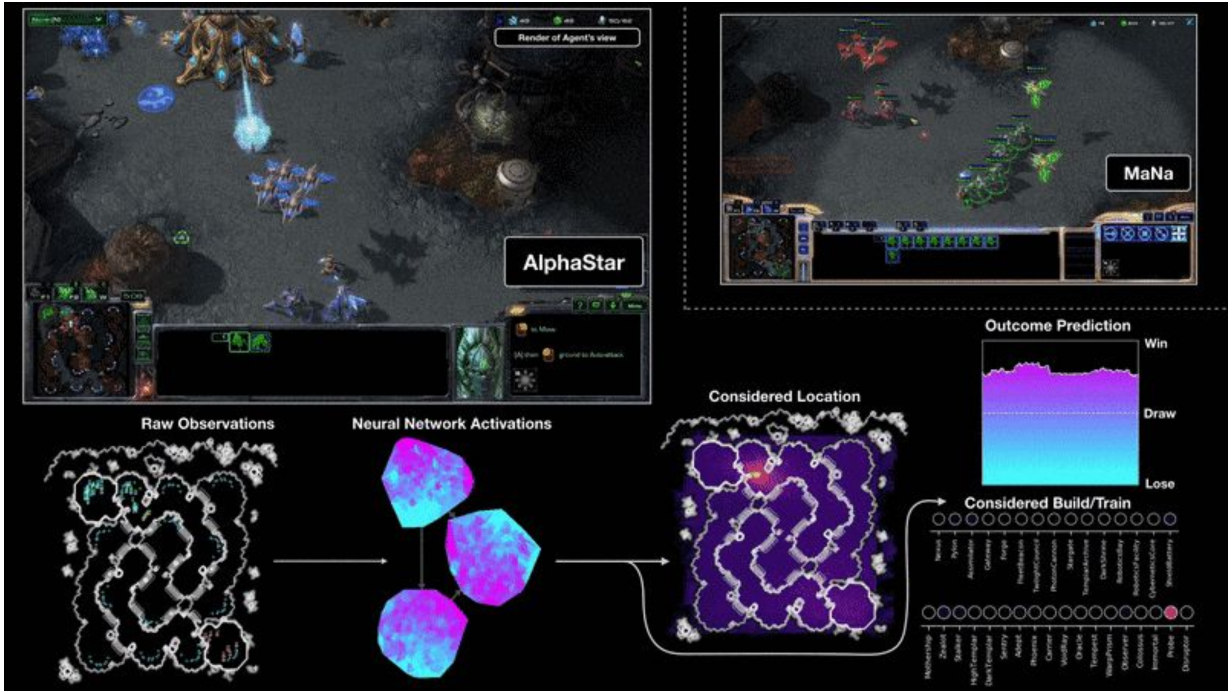
Some recent ML achievements

- DeepMind
  - AlphaZero
    - Chess/Go/Shogi
  - [WaveNet]
    - Generative model for raw Audio
    - Speech/Music



Some recent ML achievements

- DeepMind AlphaStar: Beats competitive Starcraft II players



Some recent ML achievements

- OpenAI
  - Generative Language models (GPT-2)
    - Machine translation
    - Question answering

SYSTEM PROMPT (HUMAN-WRITTEN)	*A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown.*
MODEL COMPLETION (MACHINE-WRITTEN, FIRST TRY)	The incident occurred on the downtown train line, which runs from Covington and Ashland stations. In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief. “The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation,” said Tom Hicks, the U.S. Energy Secretary, in a statement. “Our top priority is to secure the theft and ensure it doesn’t happen again.”

Hang on

- All of these results are impressive
  - Scalability of training/prediction
  - Benchmarking in real-world settings
  - Massive models, massive data-sets

- Do we have enough tools to understand these models?

Why do we need to understand models?

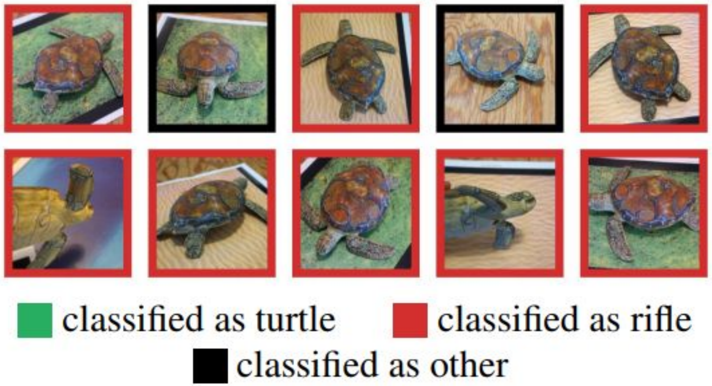
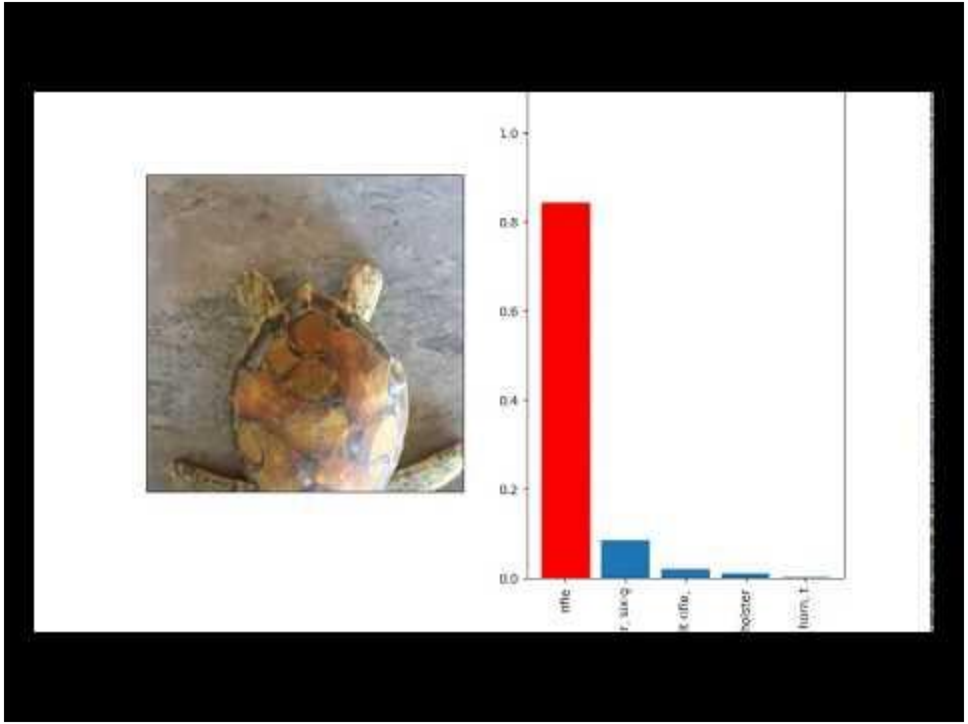


Figure 1. Randomly sampled poses of a 3D-printed turtle adversarially perturbed to classify as a rifle at every viewpoint[2]. An unperturbed model is classified correctly as a turtle nearly 100% of the time.

Athalye et al.
Synthesizing Robust Adversarial Examples
ICML 2018

Why do we need to understand models?



Athalye et al.
Synthesizing Robust Adversarial Examples
ICML 2018

Why do we need to understand models?

- Rise of ML-based systems
    - Complex, possibly interdependent black-boxes

Some ML models

- OpenAI's language model GPT-2 contains ~1.5 billion parameters

Model	Size	Top-1 Accuracy	Top-5 Accuracy	Parameters	Depth
Xception	88 MB	0.790	0.945	22,910,480	126
VGG16	528 MB	0.713	0.901	138,357,544	23
VGG19	549 MB	0.713	0.900	143,667,240	26
ResNet50	98 MB	0.749	0.921	25,636,712	-
ResNet101	171 MB	0.764	0.928	44,707,176	-
ResNet152	232 MB	0.766	0.931	60,419,944	-
ResNet50V2	98 MB	0.760	0.930	25,613,800	-
ResNet101V2	171 MB	0.772	0.938	44,675,560	-
ResNet152V2	232 MB	0.780	0.942	60,380,648	-
ResNeXt50	96 MB	0.777	0.938	25,097,128	-
ResNeXt101	170 MB	0.787	0.943	44,315,560	-
InceptionV3	92 MB	0.779	0.937	23,851,784	159
InceptionResNetV2	215 MB	0.803	0.953	55,873,736	572
MobileNet	16 MB	0.704	0.895	4,253,864	88
MobileNetV2	14 MB	0.713	0.901	3,538,984	88
DenseNet121	33 MB	0.750	0.923	8,062,504	121
DenseNet169	57 MB	0.762	0.932	14,307,880	169
DenseNet201	80 MB	0.773	0.936	20,242,984	201
NASNetMobile	23 MB	0.744	0.919	5,326,716	-
NASNetLarge	343 MB	0.825	0.960	88,949,818	-

The top-1 and top-5 accuracy refers to the model's performance on the ImageNet validation dataset.

Depth refers to the topological depth of the network. This includes activation layers, batch normalization layers etc.

https://keras.io/applications

ML Systems Today

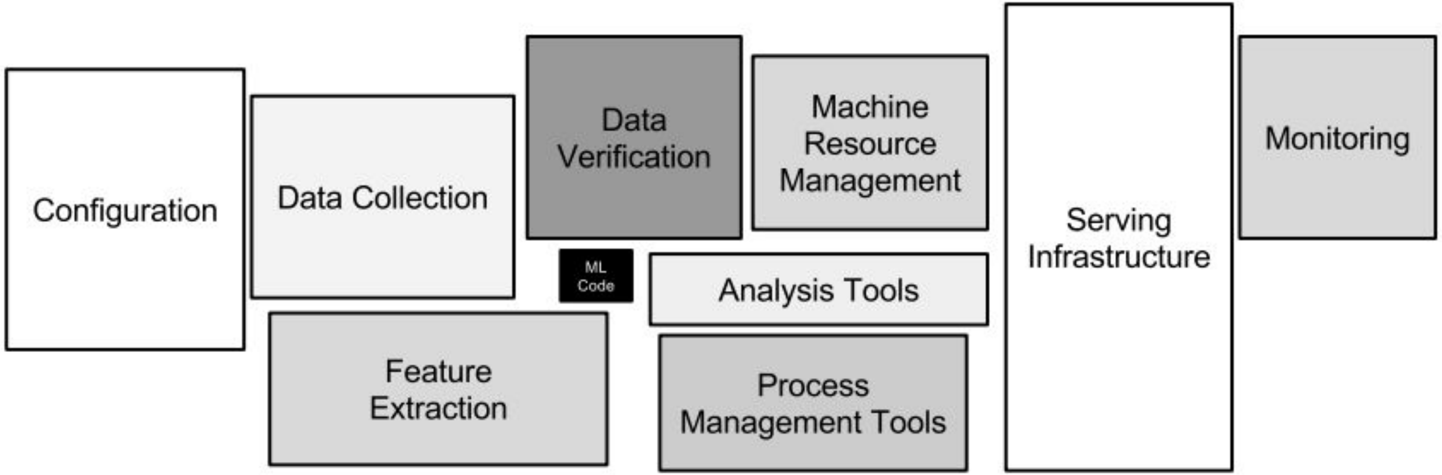


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

Hidden Technical Debt in Machine Learning Systems, NIPS 2015, Sculley et al.

Why do we need to understand models?

- Rise of ML-based systems
  - Complex, possibly interdependent black-boxes
- GDPR
  - In effect since 05/2018 in EU
  - “Meaningful explanations of the logic involved” for automated decision systems

Why do we need to understand models?

- Rise of ML-based systems
  - Complex, possibly interdependent black-boxes
- GDPR
  - In effect since 05/2018 in EU
  - “Meaningful explanations of the logic involved” for automated decision systems
- Applications
  - (Social) credit scoring
  - Health
  - Safety

Social Credit System (China)

- Government owned assessment of economic and social reputation of chinese citizens

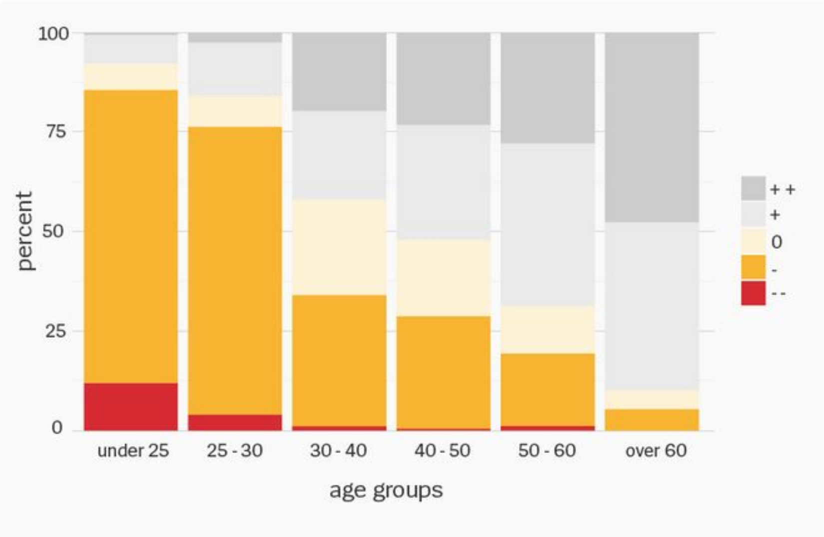
- Chinese courts banned people from buying Airplane and Train tickets more than 20 million times in 2018

Credit Scoring in Germany

- OpenSchufa
  - An initiative to reverse engineer scoring algorithm

- Schufa declared a trade secret by the Federal Court of Justice

Younger men more often have bad ratings than older ones



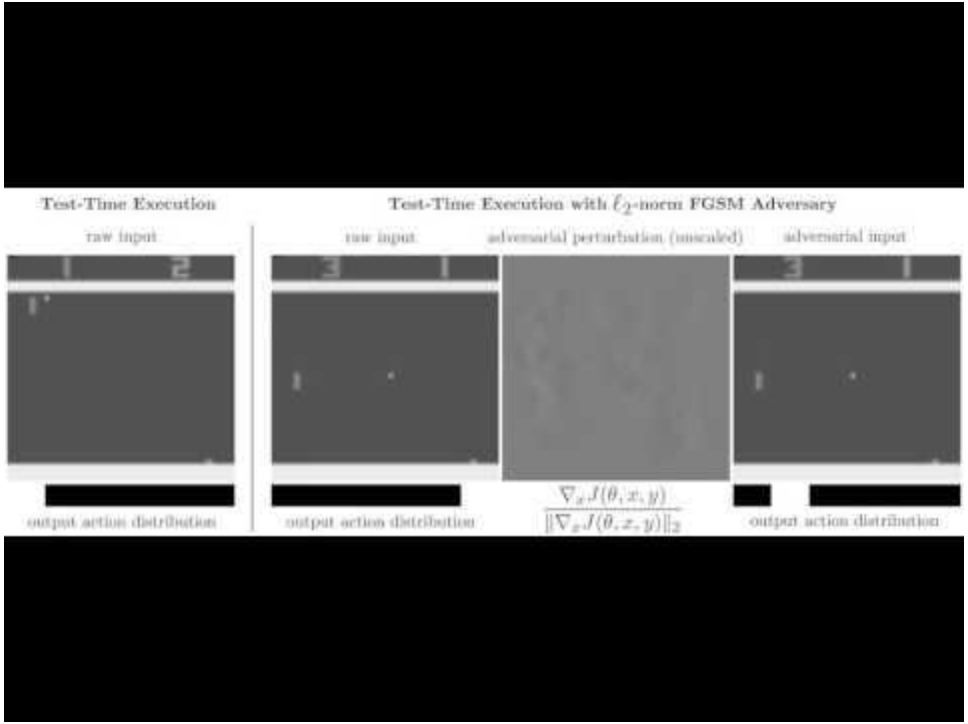
Classifications in the category „General Data“ across different age groups.
Population: all male consumers in the dataset with less than three relocations.

https://web.br.de/interaktiv/erhoehtes-risiko/english/index.html

DATA SCIENCE RETREAT®

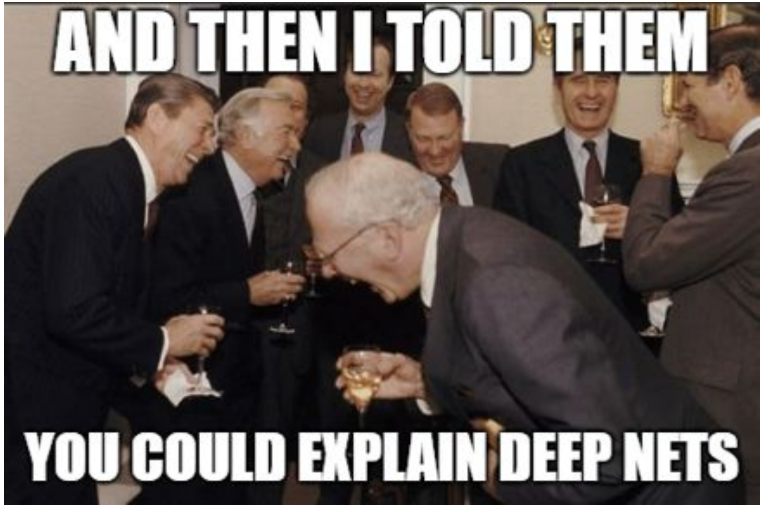
Safety

- Adversarial Attack on Deep Q-Network



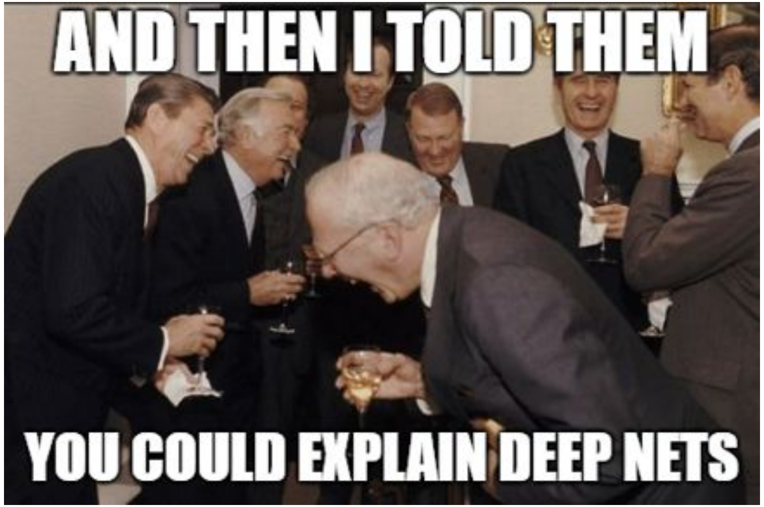
Huang, Sandy et al.
Adversarial Attacks on Neural Network Policies
arXiv preprint arXiv:1702.02284 (2017)

ML what?



ML what?

- Research is focussing on understanding and interpreting ML models
- How to define Interpretability?



DATA SCIENCE RETREAT

Interpretability Desiderata

- Trust
  - In model performance
  - …
  - Drop human from the loop?

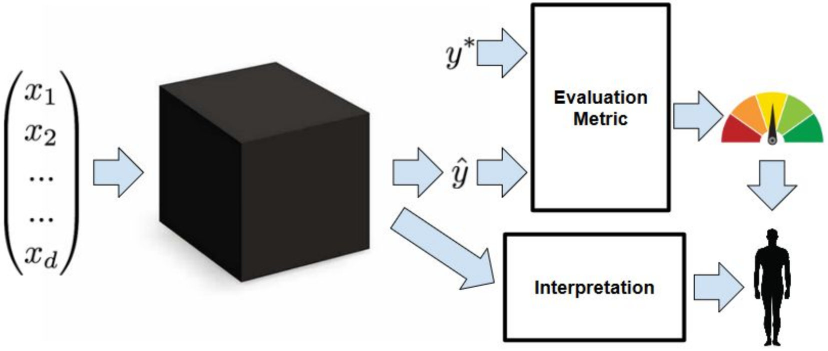


Figure 1. Typically, evaluation metrics require only predictions and *ground truth* labels. When stakeholders additionally demand *interpretability*, we might infer the existence of desiderata that cannot be captured in this fashion.

Lipton
The mythos of model interpretability
arXiv preprint arXiv:1606.03490 (2016)

Interpretability Desiderata

- Causality
  - Models learned mostly from observational data

TABLE 4.1: The results of fitting a logistic regression model on the cervical cancer dataset. Shown are the features used in the model, their estimated weights and corresponding odds ratios, and the standard errors of the estimated weights.

	Weight	Odds ratio	Std. Error
Intercept	2.91	18.36	0.32
Hormonal contraceptives y/n	0.12	1.12	0.30
Smokes y/n	-0.26	0.77	0.37
Num. of pregnancies	-0.04	0.96	0.10
Num. of diagnosed STDs	-0.82	0.44	0.33
Intrauterine device y/n	-0.62	0.54	0.40

Interpretation of a numerical feature ("Num. of diagnosed STDs"): An increase in the number of diagnosed STDs (sexually transmitted diseases) changes (decreases) the odds of cancer vs. no cancer by a factor of 0.44, when all other features remain the same. Keep in mind that correlation does not imply causation. No recommendation here to get STDs.

Molnar
Interpretable Machine Learning
leanpub

Interpretability Desiderata

- Fairness
  - Is the model fair?

- Can fairness be measured objectively?
  - Dozens of metrics for fairness in social sciences and statistics
    - Demographic parity
    - Group thresholds
    - ...

https://pair-code.github.io/what-if-tool/ai-fairness.html

Interpretability Desiderata

- What do you think?

Paper session

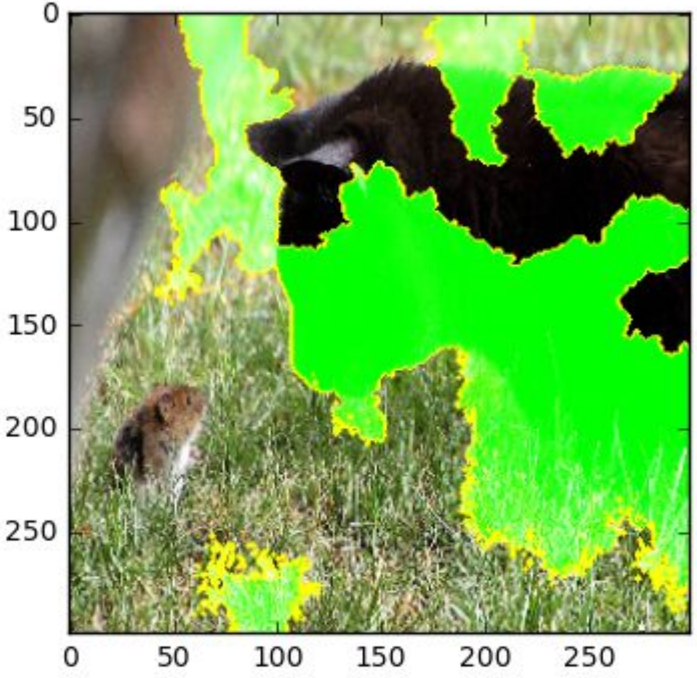
- Read a paper
  - Position paper: Mythos
  - Technical paper: Lapuschkin
- Discussion

Getting Started

- Clone GitHub repo
  - **git clone git@github.com:tdhd/interpretability-class.git**
- Setup local environment
  - Python packages
  - Docker
- Questions?

LIME

- Explain black box predictors
- Vision and Text-based models



LIME

- How does it work?
  - Select instance of interest for which you want to have an explanation
  - Perturb your dataset and get the black box predictions for these new points
  - Weight the new samples according to their proximity to the instance of interest
  - Train a weighted, interpretable model on the dataset with the variations
  - Explain the prediction by interpreting the local model
- Depending on input domain, pertubation different
  - Text: Randomly remove words
  - Images: Randomly remove super-pixels

DATA SCIENCE RETREAT

ELI5

- Explain and Debug ML models (scikit-learn/XGBoost/...)
- Provides rich visualizations based on LIME/Permutation Importance

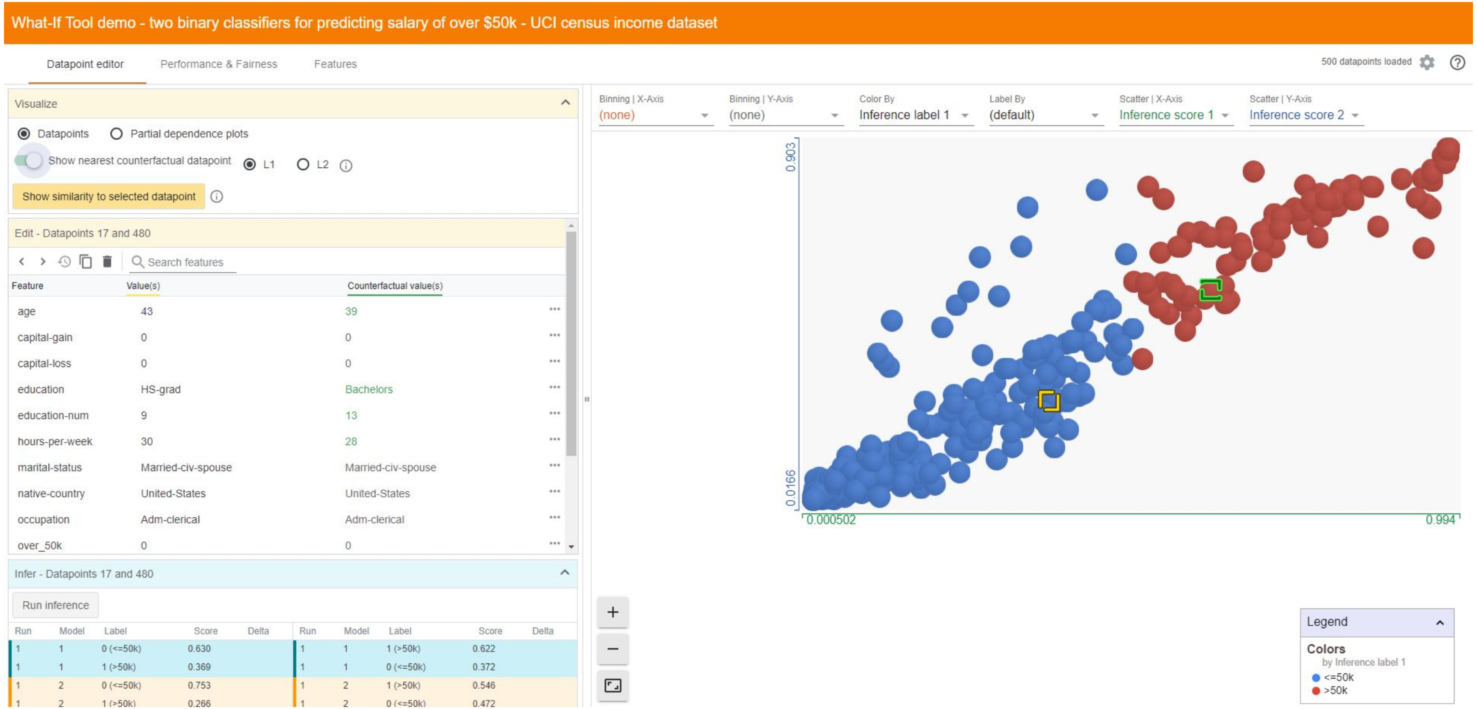
y=alt.atheism top features		**y=comp.graphics** top features		**y=sci.med** top features		**y=soc.religion.christian** top features	
Weight?	**Feature**	**Weight?**	**Feature**	**Weight?**	**Feature**	**Weight?**	**Feature**
+1.991	mathew	+1.702	graphics	+2.016	pitt	+1.193	rutgers
+1.925	keith	+0.825	images	+1.951	doctor	+1.030	church
+1.834	atheism	+0.798	files	+1.758	information	+1.021	christians
+1.813	okcforum	+0.786	software	+1.697	disease	+0.946	clh
+1.697	go	+0.779	file	+1.655	treatment	+0.899	christ
+1.696	psuvm	+0.773	image	+1.522	msg	+0.797	christian
+1.617	believing	+0.729	package	+1.518	health	*... 11122 more positive ...*	
+1.594	psu	+0.724	card	*... 15007 more positive ...*		*... 24657 more negative ...*	
... 10174 more positive ...		+0.702	3d	*... 20772 more negative ...*		-0.852	graphics
... 25605 more negative ...		*... 11710 more positive ...*		-1.764	god	-0.894	posting
-1.686	rutgers	*... 24069 more negative ...*		-2.171	graphics	-1.181	nntp
-10.453	<BIAS>	-1.379	<BIAS>	-5.013	<BIAS>	-1.243	host

What-If Tool

- Dataset and model exploration

What-If Tool

- Counterfactuals



Case study

- Explain model(s) with Tool of choice, e.g.,
  - What-If-Tool
  - ELI5
  - iNNvestigate
  - shap
  - ...
- Ideas for which models to explain
  - From portfolio project
  - Official or Research Tensorflow models (https://github.com/tensorflow/models)
  - BERT/Logreg/GPT*
  - ...

DATA SCIENCE RETREAT®