Что у ней внутри:

Зачем и как интерпретировать ML-модели

Дмитрий Колодезев, ООО Промсофт

«Семинар Невозможного», 2021.09.23

План

- Машинное обучение — оно везде
- Что это, кстати, такое
- Как делить выигрыш в карты
- Чем больше отчетов, тем хуже
- Как понять нейронную сеть
- Где интересно
- Как быть

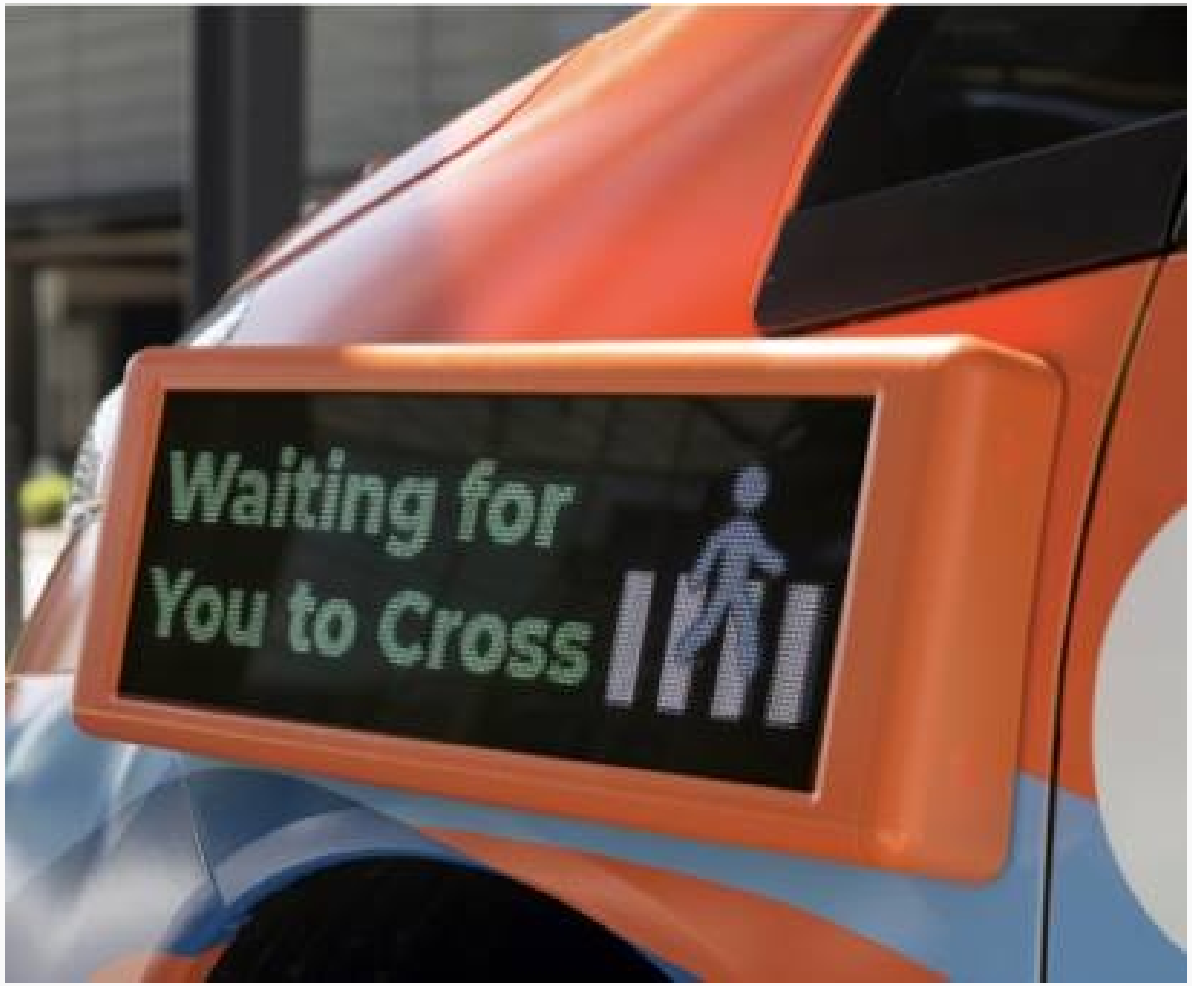
История вопроса

— Вот, изволите видеть, так называемая эвристическая машина, — сказал старичок.

— Точный электронно-механический прибор для отвечания на любые вопросы, а именно на научные и хозяйственные.

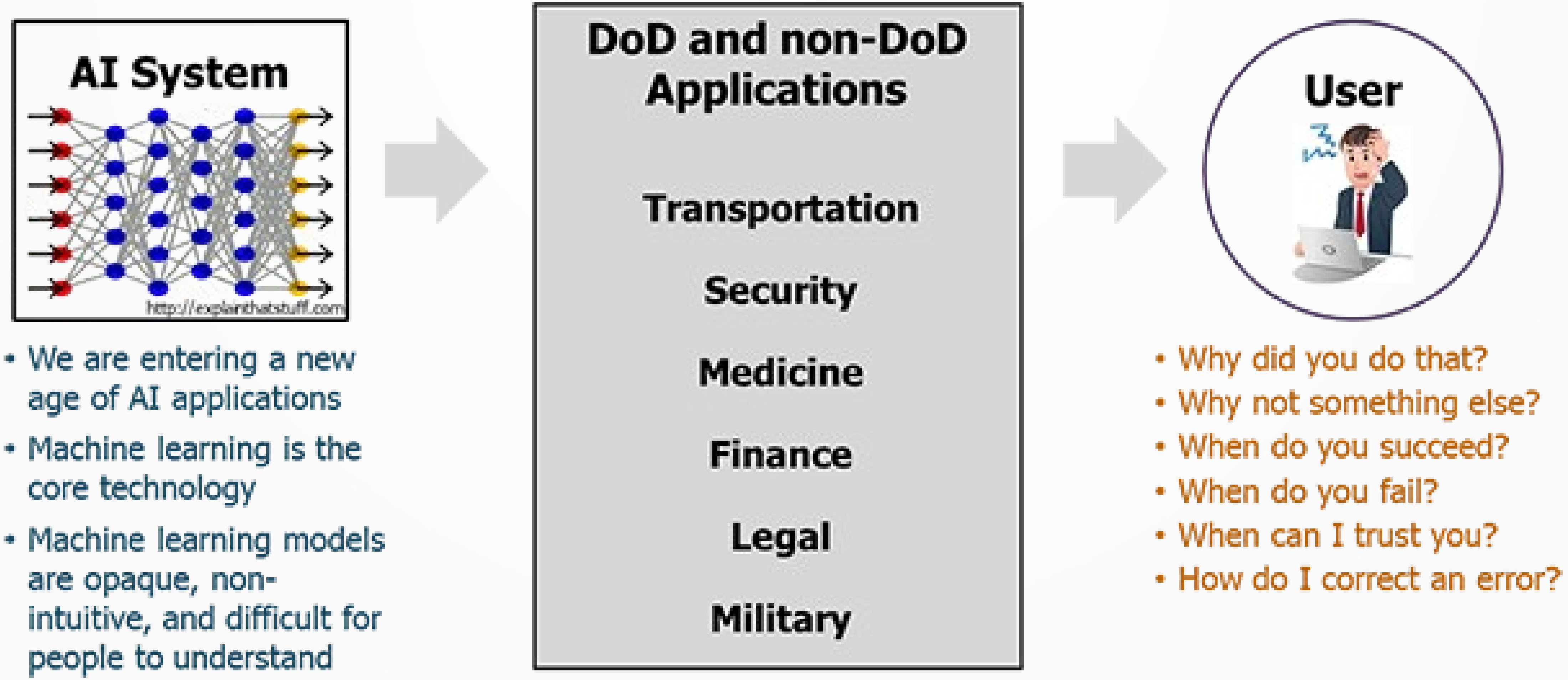
А.Б. Стругацкие, «Сказка о Тройке»

Осторожно, роботы!


Waiting for
You to Cross

Осторожно, люди

DARPA, eXplainable AI



AI System

- We are entering a new age of AI applications
- Machine learning is the core technology
- Machine learning models are opaque, non-intuitive, and difficult for people to understand

DoD and non-DoD Applications

Transportation

Security

Medicine

Finance

Legal

Military

User

- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

Explainable Artificial Intelligence (XAI)

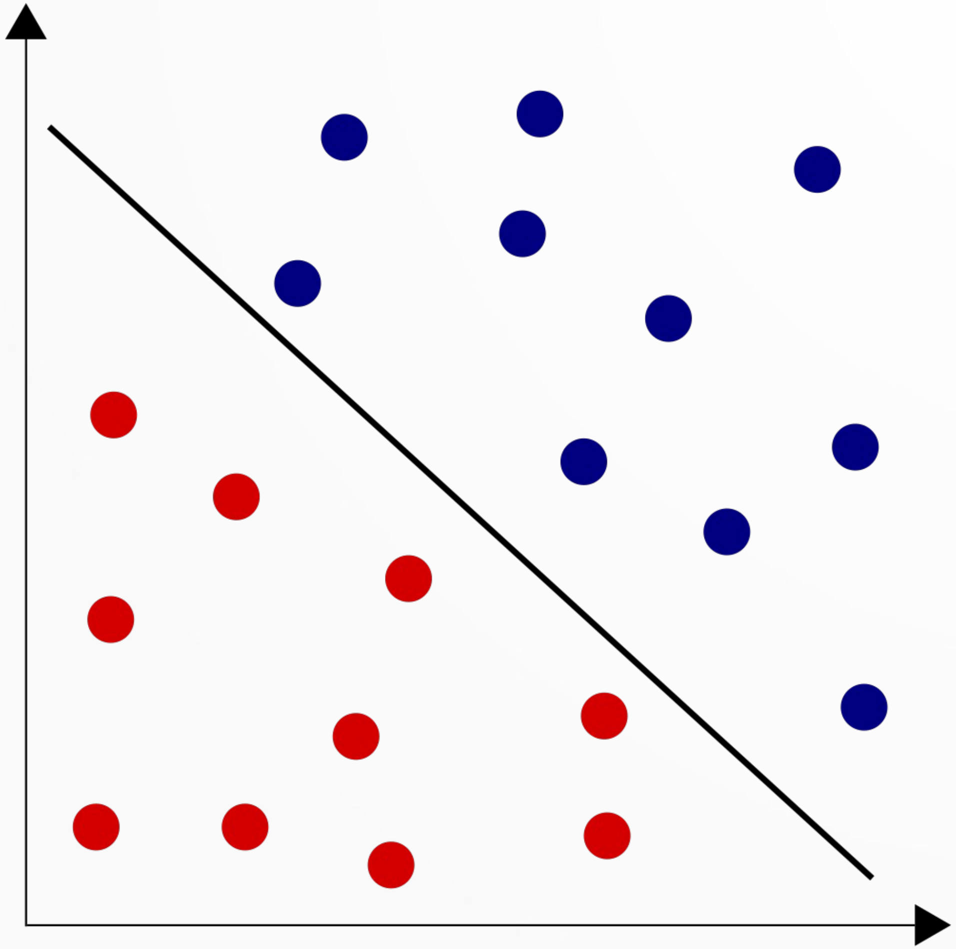
Кстати, что такое ML?

Machine learning is the science of getting computers to act without being explicitly programmed *

Software 2.0 is code written by the optimization based on an evaluation criterion (such as “classify this training data correctly”) *

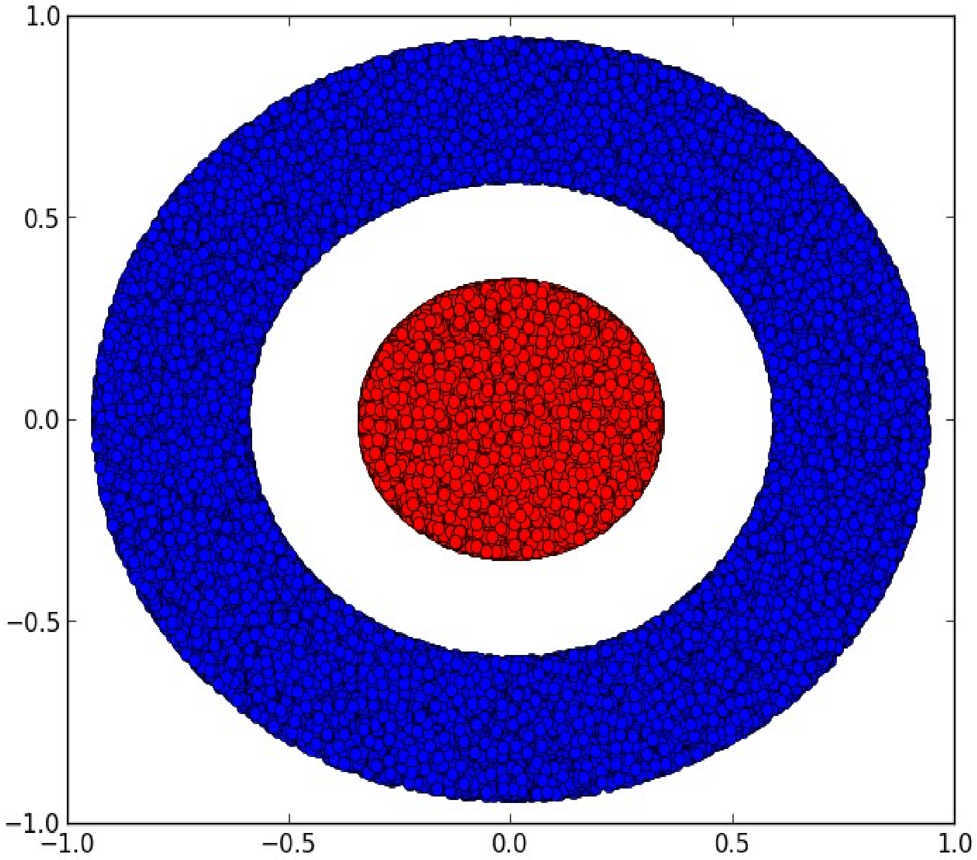
Машинное обучение - автоматизированное построение удобной нам проекции признакового пространства

Линейно разделяй и в.



- Это удобное пространство
- Линейно разделимое
- Чуток повернуть, сдвинуть, сплющить
- И вообще будет огонь

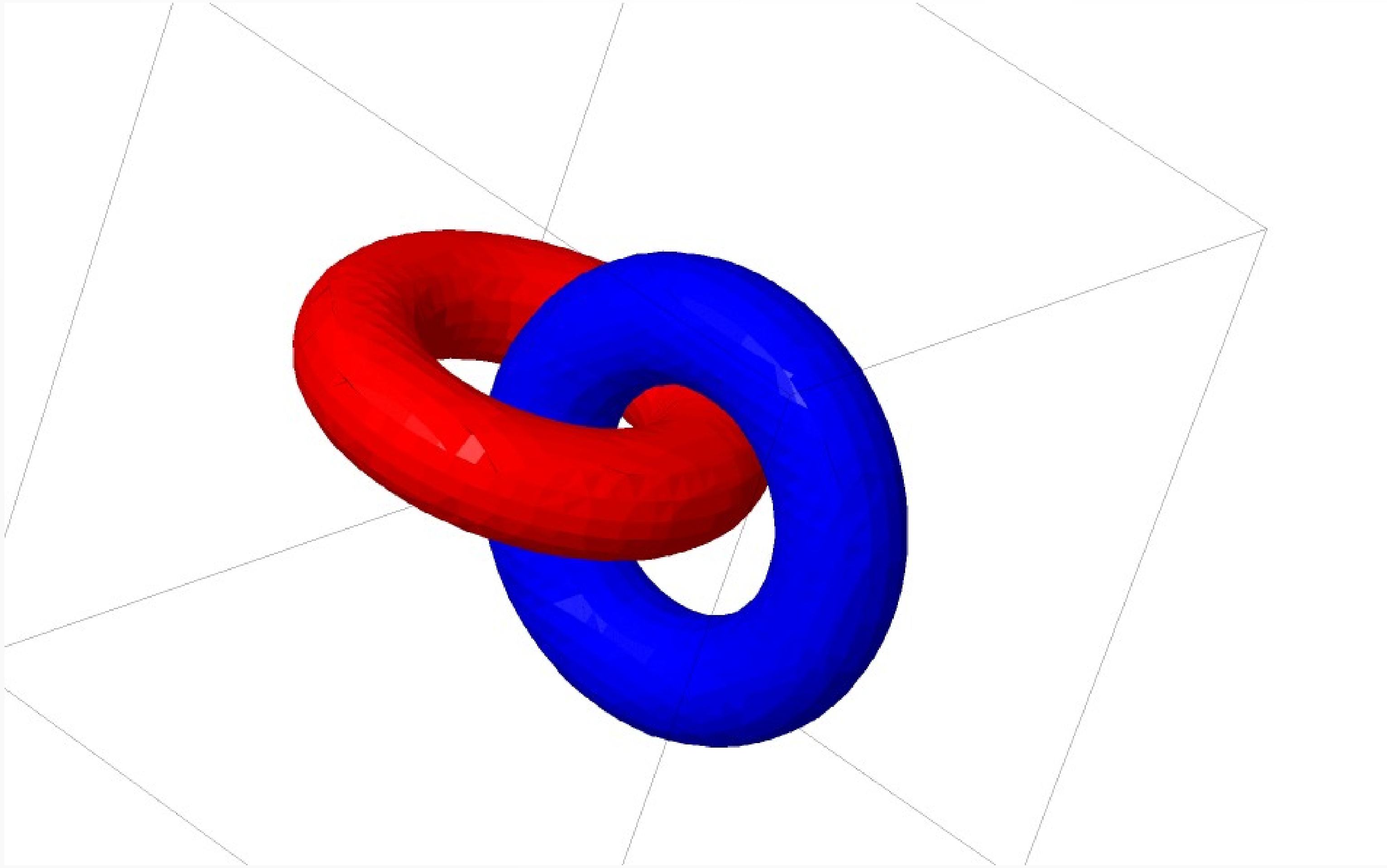
Kernel Trick



- Сколько ни крути
- А давайте перейдем в полярные координаты?

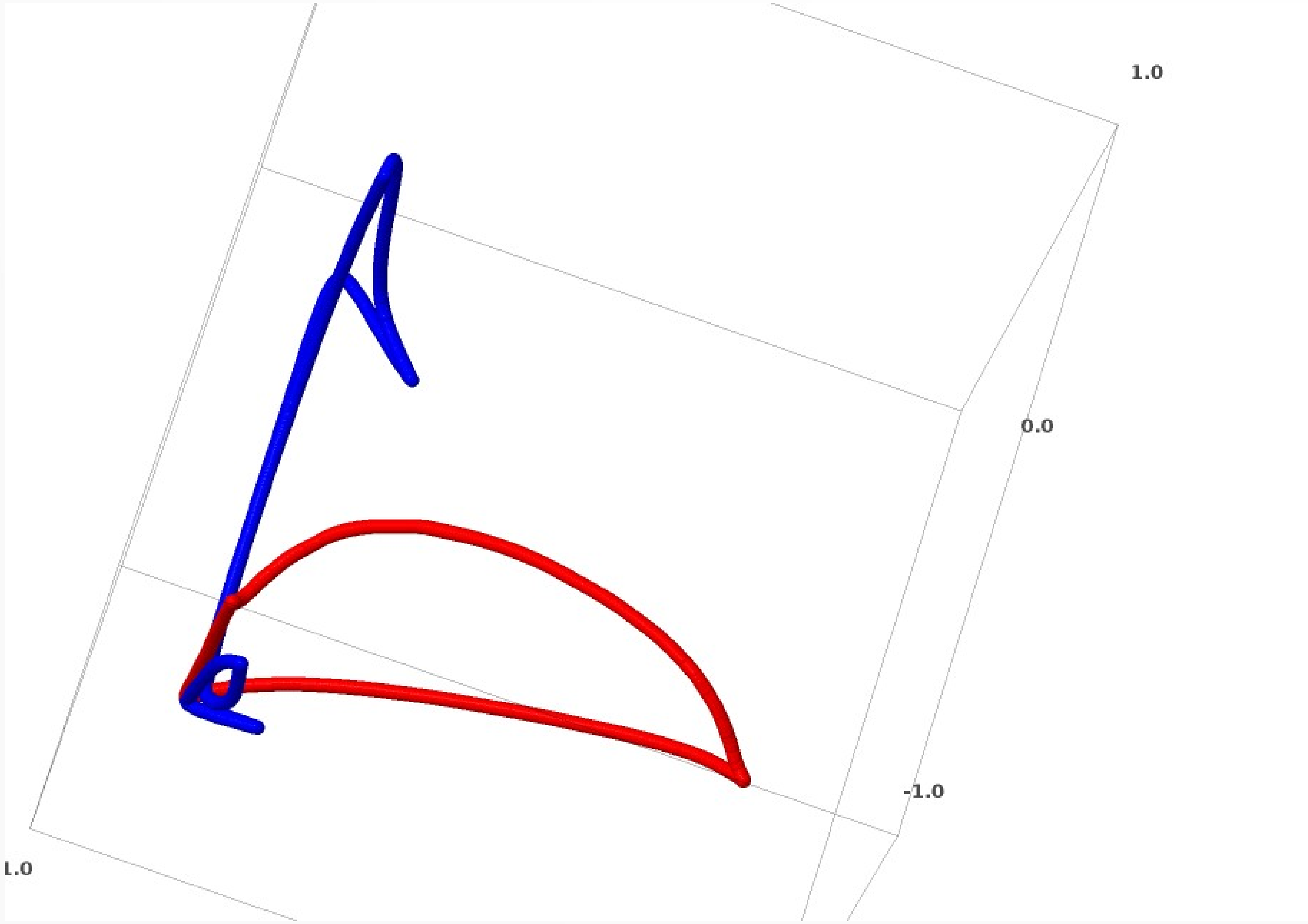
http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/

А что если



http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/

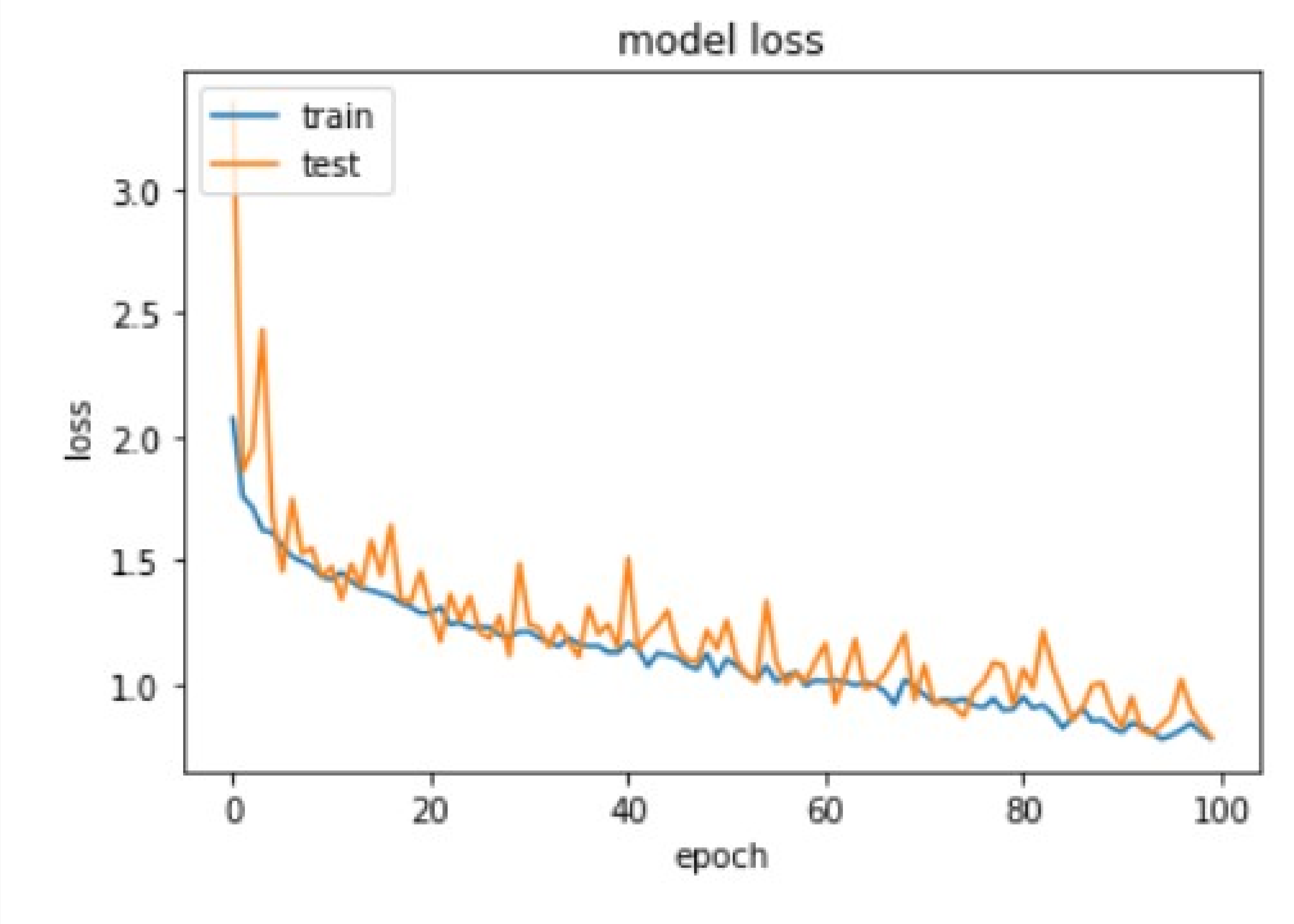
Тяни сильнее



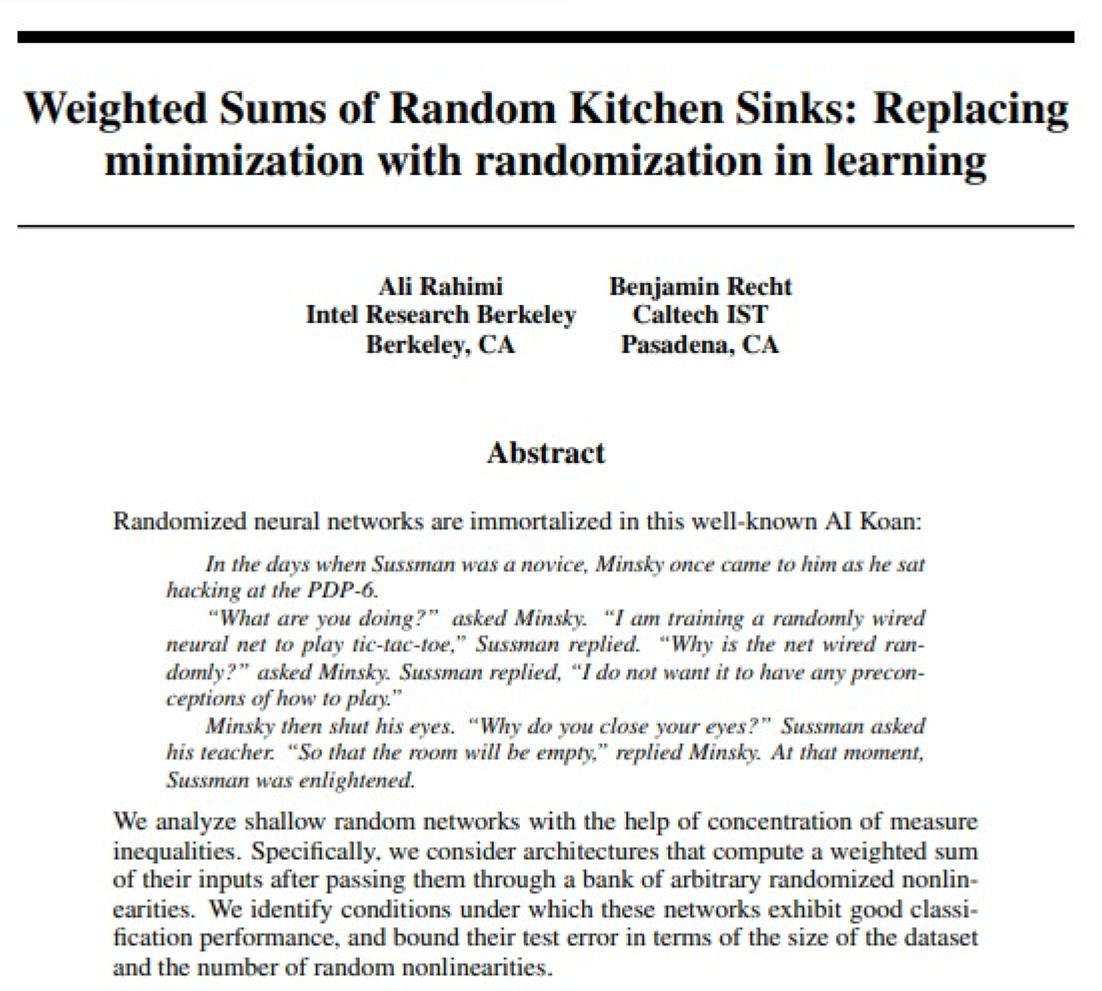
- Руби через узел
- Нейронки так и делают
- Почти разделимо
- Кому-то не повезло
- Для практики достаточно

http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/

Тянем бублики наощупь



Как-попало-проекция



Weighted Sums of Random Kitchen Sinks: Replacing minimization with randomization in learning

Ali Rahimi
Intel Research Berkeley
Berkeley, CA

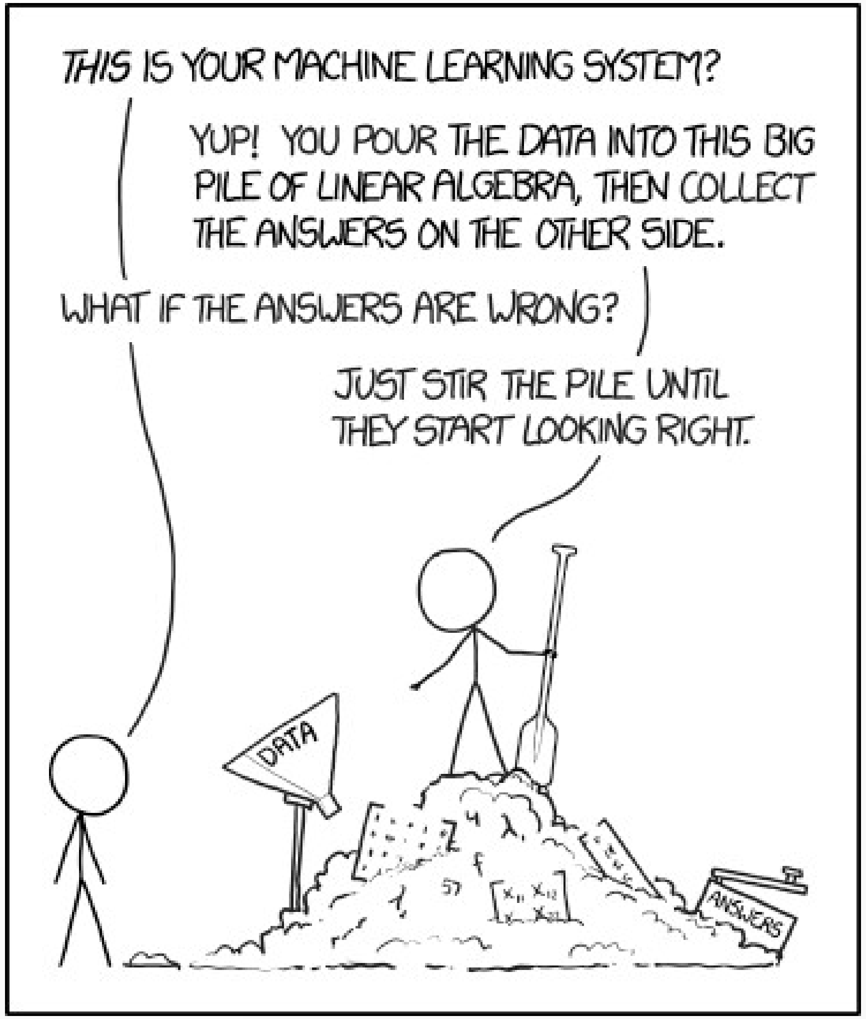
Benjamin Recht
Caltech IST
Pasadena, CA

Abstract

Randomized neural networks are immortalized in this well-known AI Koan:

> *In the days when Sussman was a novice, Minsky once came to him as he sat hacking at the PDP-6.*
>
> *"What are you doing?" asked Minsky. "I am training a randomly wired neural net to play tic-tac-toe," Sussman replied. "Why is the net wired randomly?" asked Minsky. Sussman replied, "I do not want it to have any preconceptions of how to play."*
>
> *Minsky then shut his eyes. "Why do you close your eyes?" Sussman asked his teacher. "So that the room will be empty," replied Minsky. At that moment, Sussman was enlightened.*

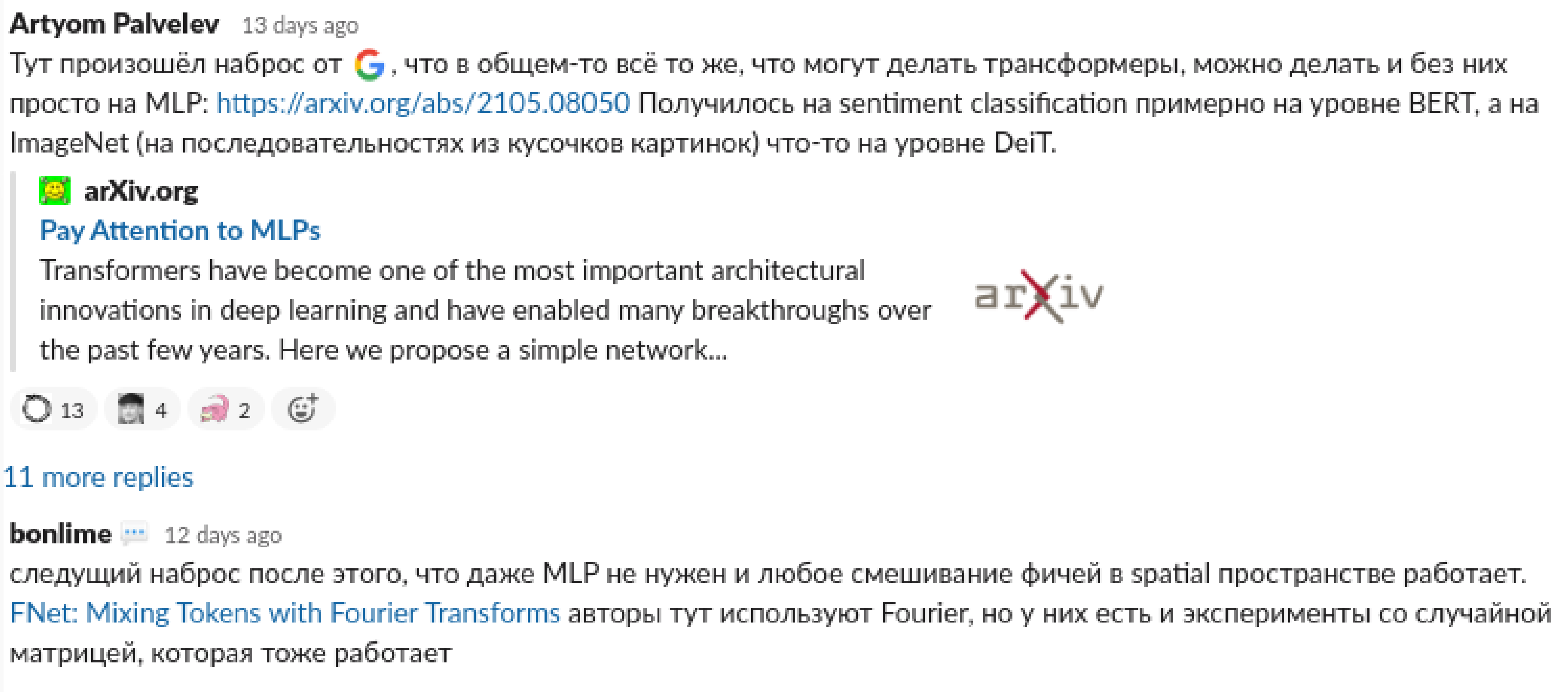
We analyze shallow random networks with the help of concentration of measure inequalities. Specifically, we consider architectures that compute a weighted sum of their inputs after passing them through a bank of arbitrary randomized nonlinearities. We identify conditions under which these networks exhibit good classification performance, and bound their test error in terms of the size of the dataset and the number of random nonlinearities.


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

https://www.youtube.com/watch?v=x7psGHgatGM

https://people.eecs.berkeley.edu/~brecht/papers/08.rah.rec.nips.pdf

Да не угаснет пламя



Artyom Palvelev 13 days ago

Тут произошёл наброс от G, что в общем-то всё то же, что могут делать трансформеры, можно делать и без них просто на MLP: https://arxiv.org/abs/2105.08050 Получилось на sentiment classification примерно на уровне BERT, а на ImageNet (на последовательностях из кусочков картинок) что-то на уровне DeiT.

> **arXiv.org**
> Pay Attention to MLPs
> Transformers have become one of the most important architectural innovations in deep learning and have enabled many breakthroughs over the past few years. Here we propose a simple network...

bonlime 12 days ago

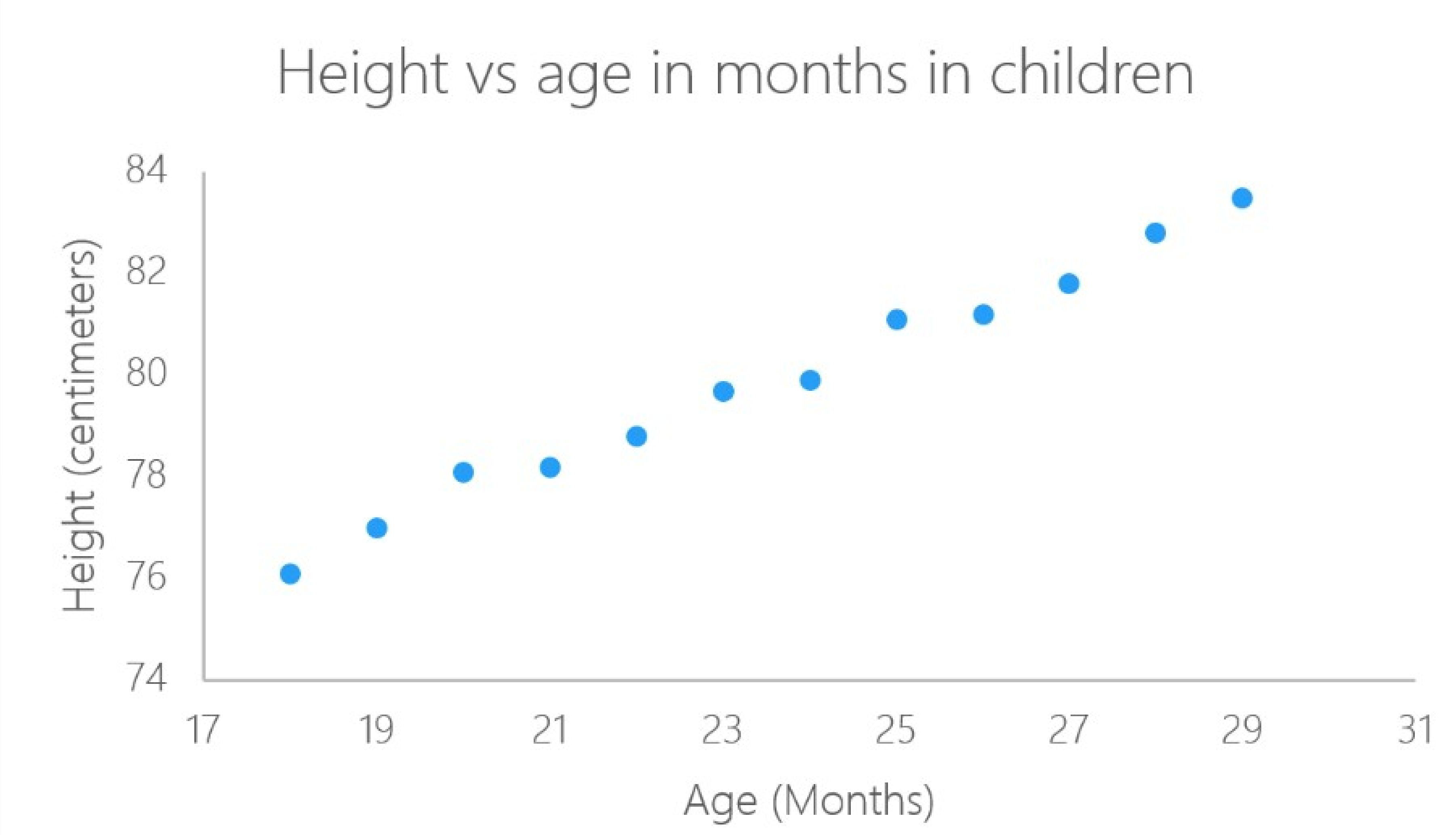
следущий наброс после этого, что даже MLP не нужен и любое смешивание фичей в spatial пространстве работает. FNet: Mixing Tokens with Fourier Transforms авторы тут используют Fourier, но у них есть и эксперименты со случайной матрицей, которая тоже работает

https://arxiv.org/abs/2105.08050 https://arxiv.org/abs/2105.03824

Как в жизни?

- Гипотеза о многообразии
- https://deepai.org/machine-learning-glossary-and-terms/manifold-hypothesis
- http://www.mit.edu/~mitter/publications/121_Testing_Manifold.pdf

Гипотеза о многообразии



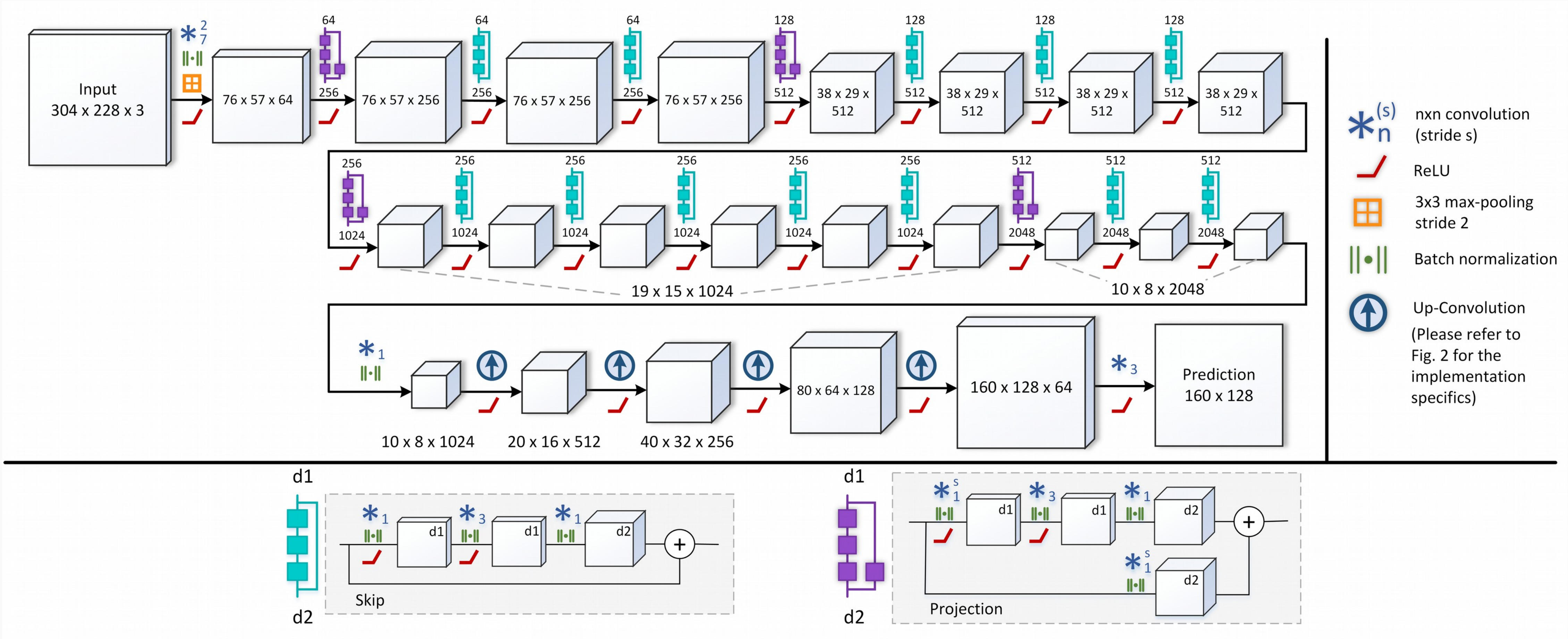
https://www.datacamp.com/community/tutorials/linear-regression-R

Итого

ML-алгоритмы строят проекции:

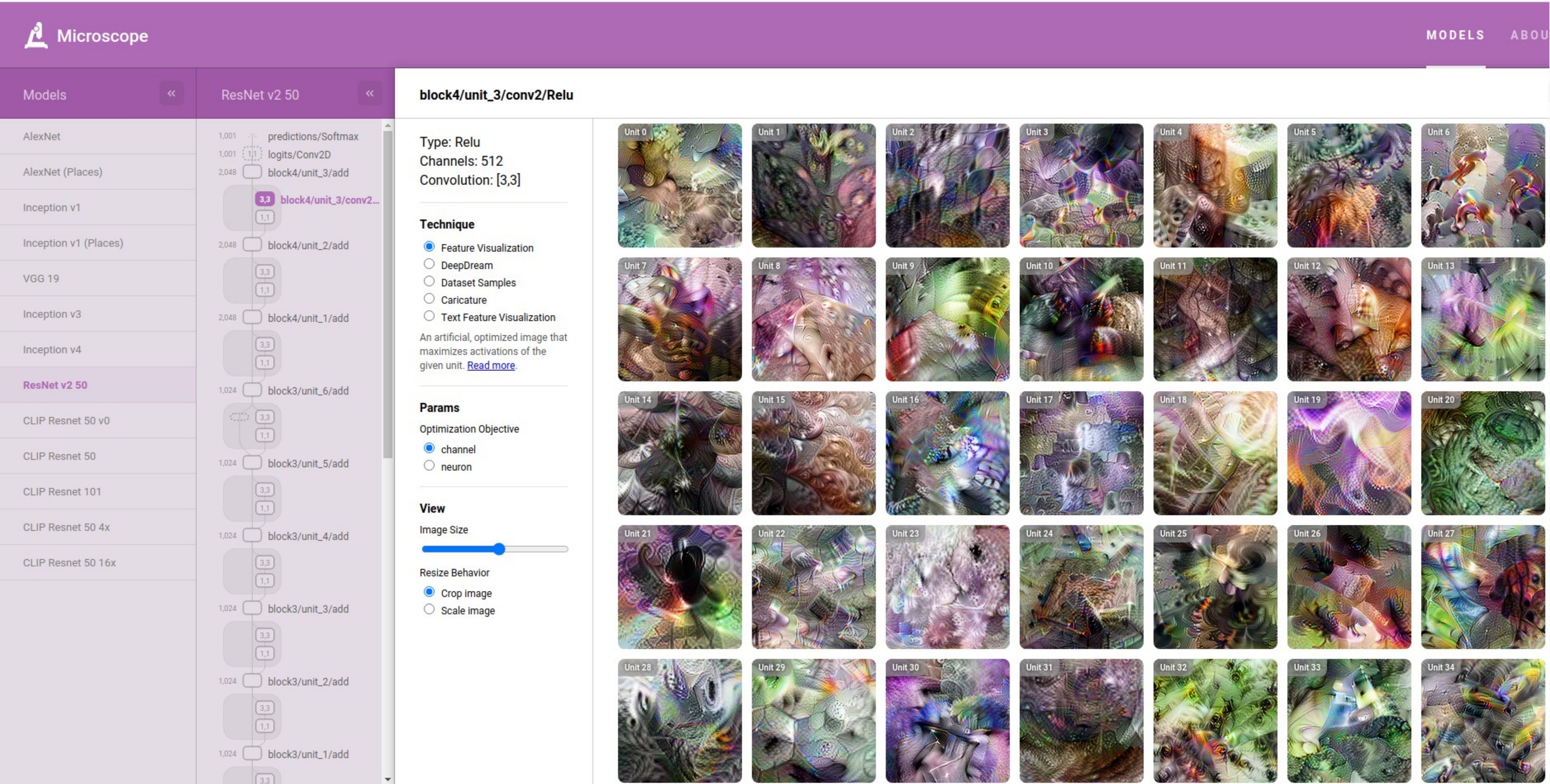
- Бесконтрольно
- Удобные им
- Нелогичные
- Нелинейные
- Немыслимые
- Технически при этом никакой магии нет

В принципе все понятно


Input
304 x 228 x 3
2
7
76 x 57 x 64
64
256
76 x 57 x 256
64
256
76 x 57 x 256
64
256
76 x 57 x 256
128
512
38 x 29 x 512
128
512
38 x 29 x 512
128
512
38 x 29 x 512
128
512
38 x 29 x 512
256
1024
256
1024
256
1024
256
1024
256
1024
256
1024
512
2048
512
2048
512
2048
19 x 15 x 1024
10 x 8 x 2048
1
10 x 8 x 1024
20 x 16 x 512
40 x 32 x 256
80 x 64 x 128
160 x 128 x 64
3
Prediction
160 x 128
(s)
n
nxn convolution
(stride s)
ReLU
3x3 max-pooling
stride 2
Batch normalization
Up-Convolution
(Please refer to
Fig. 2 for the
implementation
specifics)
d1
d2
1
3
1
d1
d1
d2
+
Skip
d1
d2
s
1
3
1
s
1
d1
d1
d2
d2
+
Projection

https://www.arxiv-vanity.com/papers/1606.00373/

OpenAI Microscope


Microscope
MODELS
ABOU
Models
ResNet v2 50
AlexNet
AlexNet (Places)
Inception v1
Inception v1 (Places)
VGG 19
Inception v3
Inception v4
ResNet v2 50
CLIP Resnet 50 v0
CLIP Resnet 50
CLIP Resnet 101
CLIP Resnet 50 4x
CLIP Resnet 50 16x
1,001 predictions/Softmax
1,001 logits/Conv2D
2,048 block4/unit_3/add
block4/unit_3/conv2...
2,048 block4/unit_2/add
2,048 block4/unit_1/add
1,024 block3/unit_6/add
1,024 block3/unit_5/add
1,024 block3/unit_4/add
1,024 block3/unit_3/add
1,024 block3/unit_2/add
1,024 block3/unit_1/add
Unit 0
Unit 1
Unit 2
Unit 3
Unit 4
Unit 5
Unit 6
Unit 7
Unit 8
Unit 9
Unit 10
Unit 11
Unit 12
Unit 13
Unit 14
Unit 15
Unit 16
Unit 17
Unit 18
Unit 19
Unit 20
Unit 21
Unit 22
Unit 23
Unit 24
Unit 25
Unit 26
Unit 27
Unit 28
Unit 29
Unit 30
Unit 31
Unit 32
Unit 33
Unit 34

block4/unit_3/conv2/Relu

Type: Relu
Channels: 512
Convolution: [3,3]

Technique

- ● Feature Visualization
- ○ DeepDream
- ○ Dataset Samples
- ○ Caricature
- ○ Text Feature Visualization

An artificial, optimized image that maximizes activations of the given unit. Read more.

Params

Optimization Objective

- ● channel
- ○ neuron

View

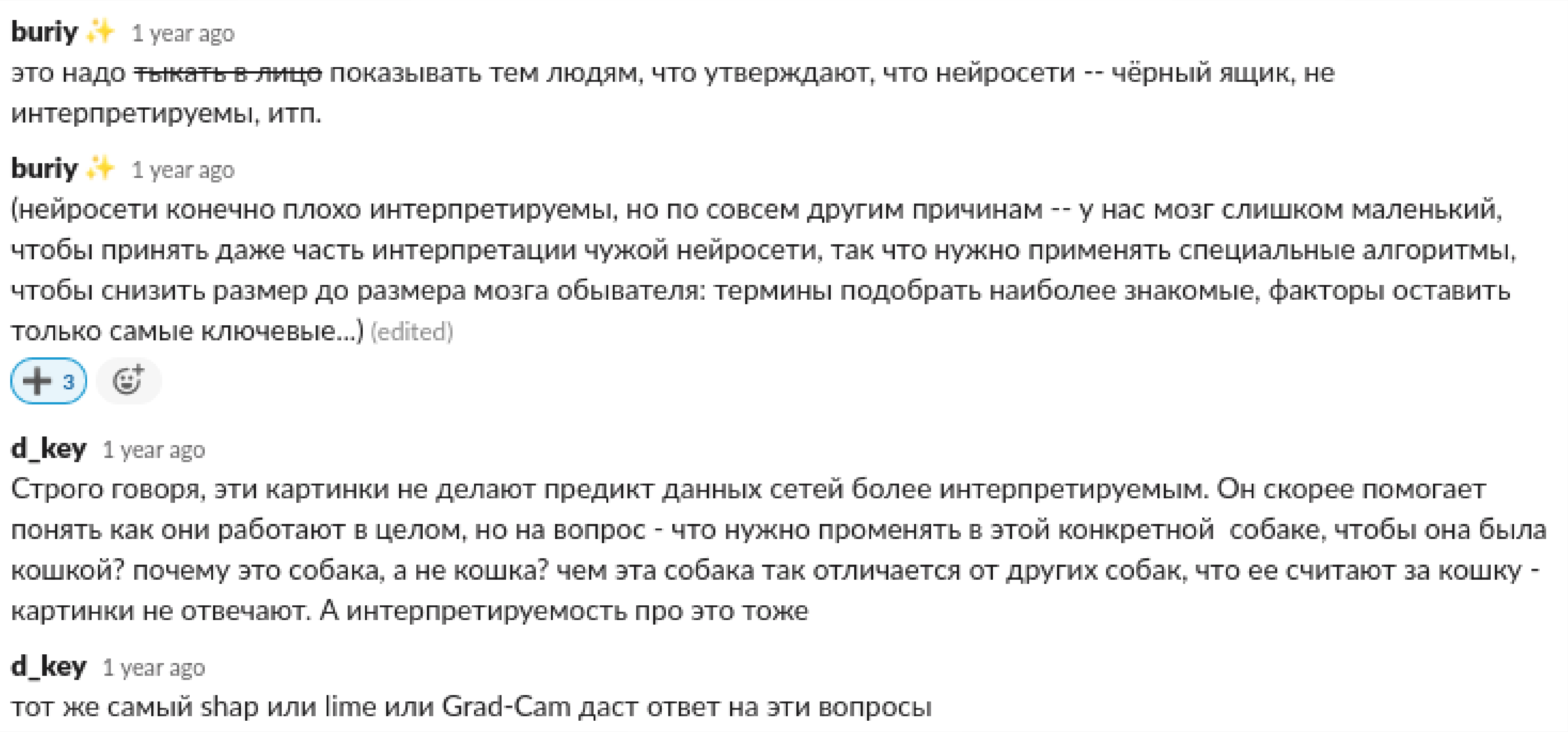
Image Size

Resize Behavior

- ● Crop image
- ○ Scale image

Проблема не в моделях

https://openai.com/blog/microscope/



buriy ✨ 1 year ago

это надо ~~тыкать в лицо~~ показывать тем людям, что утверждают, что нейросети -- чёрный ящик, не интерпретируемы, итп.

buriy ✨ 1 year ago

(нейросети конечно плохо интерпретируемы, но по совсем другим причинам -- у нас мозг слишком маленький, чтобы принять даже часть интерпретации чужой нейросети, так что нужно применять специальные алгоритмы, чтобы снизить размер до размера мозга обывателя: термины подобрать наиболее знакомые, факторы оставить только самые ключевые...) (edited)

d_key 1 year ago

Строго говоря, эти картинки не делают предикт данных сетей более интерпретируемым. Он скорее помогает понять как они работают в целом, но на вопрос - что нужно променять в этой конкретной собаке, чтобы она была кошкой? почему это собака, а не кошка? чем эта собака так отличается от других собак, что ее считают за кошку - картинки не отвечают. А интерпретируемость про это тоже

d_key 1 year ago

тот же самый shap или lime или Grad-Cam даст ответ на эти вопросы

Но проблема есть

- Разработчик. Подсказки:
  - Почему она здесь ошиблась?
  - Какие примеры научили ее плохому?
- Пользователь. Предсказуемость:
  - Чего от нее ждать?
  - Границы применимости?
- Следователь. Соответствие требованиям:
  - Почему она сделала это?
  - Кто за это ответит?

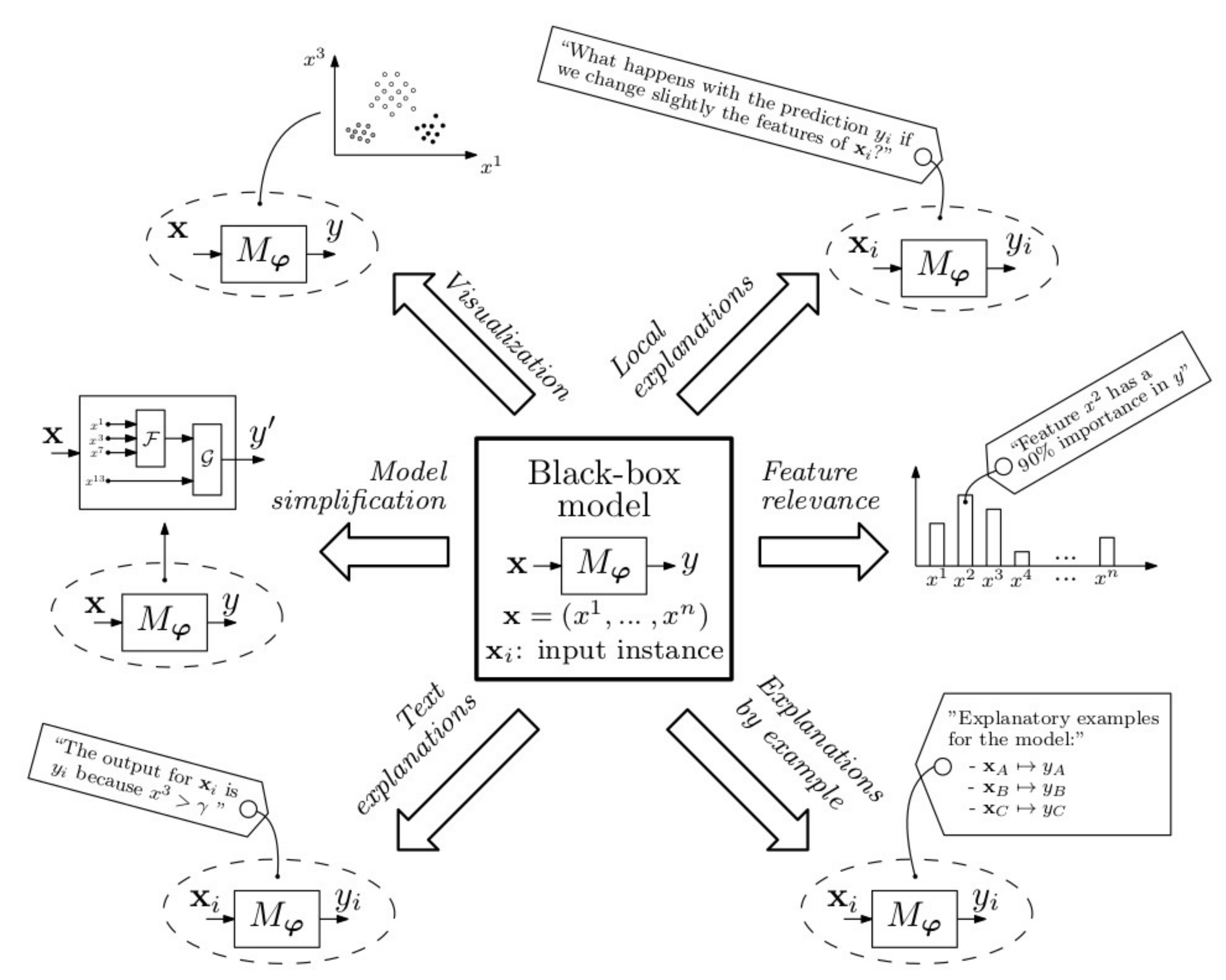
Возможно ли это?

— «У мене внутре… гм… не… неонка»

— Внутре смотрите,
где у нее анализатор и думатель..

А.Б. Стругацкие, «Сказка о Тройке»

Подходы


x^3
x^1
"What happens with the prediction y_i if we change slightly the features of $\mathbf{x}_i$?"
$\mathbf{x}$ M_φ y
Visualization
Local explanations
$\mathbf{x}_i$ M_φ y_i
$\mathbf{x}$
x^1 x^3 x^7 x^{13}
$\mathcal{F}$ $\mathcal{G}$
y'
Model simplification
Black-box model
$\mathbf{x}$ M_φ y
$\mathbf{x} = (x^1, \dots, x^n)$
$\mathbf{x}_i$: input instance
Feature relevance
"Feature x^2 has a 90% importance in y"
x^1 x^2 x^3 x^4 … x^n
$\mathbf{x}$ M_φ y
Text explanations
Explanations by example
"The output for $\mathbf{x}_i$ is y_i because $x^3 > \gamma$"
"Explanatory examples for the model:"
- $\mathbf{x}_A \mapsto y_A$
- $\mathbf{x}_B \mapsto y_B$
- $\mathbf{x}_C \mapsto y_C$
$\mathbf{x}_i$ M_φ y_i
$\mathbf{x}_i$ M_φ y_i

https://arxiv.org/abs/1910.10045

- Визуализация
- Упрощение
- Локальные объяснения
- Важность признаков
- Объяснение на примерах
- Текстовые объяснения

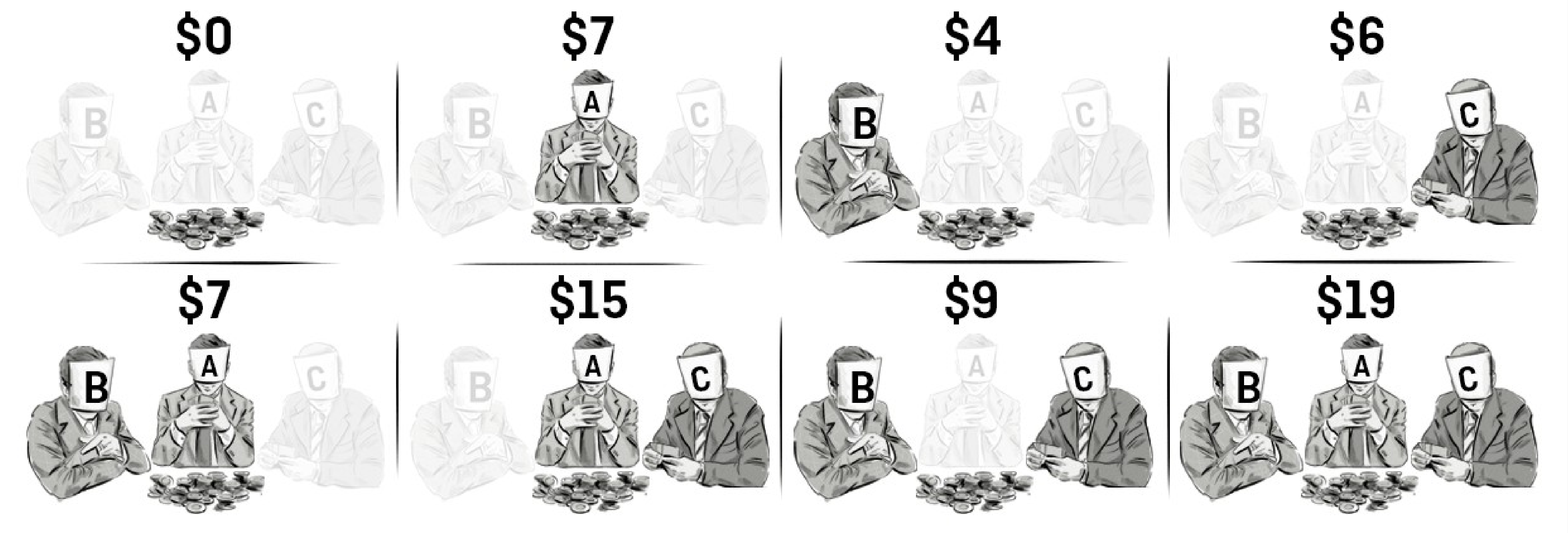
а еще

- Влияние обучающих примеров
- Граф взаимодействия

Варианты

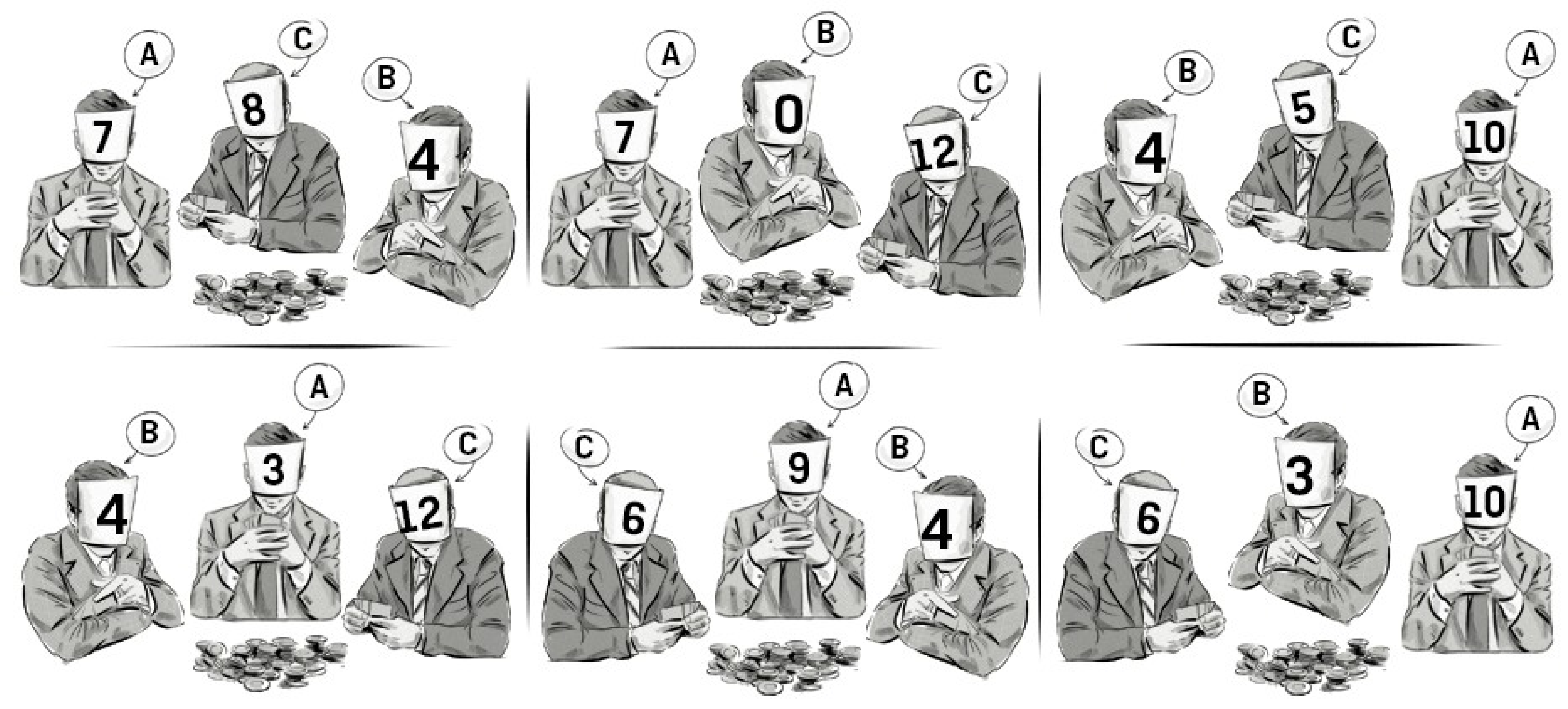
- Суррогатные модели
  - Shap, Grad-CAM
- Ближайшие соседи
- Прототипы и критики
  - MMD-critic, обычно вручную
- Якоря
  - Alibi
- Тысячи их

Shapley Values


$0
B
A
C
$7
B
A
C
$4
B
A
C
$6
B
A
C
$7
B
A
C
$15
B
A
C
$9
B
A
C
$19
B
A
C

https://clearcode.cc/blog/game-theory-attribution/

Shapley Values


A
7
C
8
B
4
A
7
B
0
C
12
B
4
C
5
A
10
B
4
A
3
C
12
C
6
A
9
B
4
C
6
B
3
A
10

https://clearcode.cc/blog/game-theory-attribution/

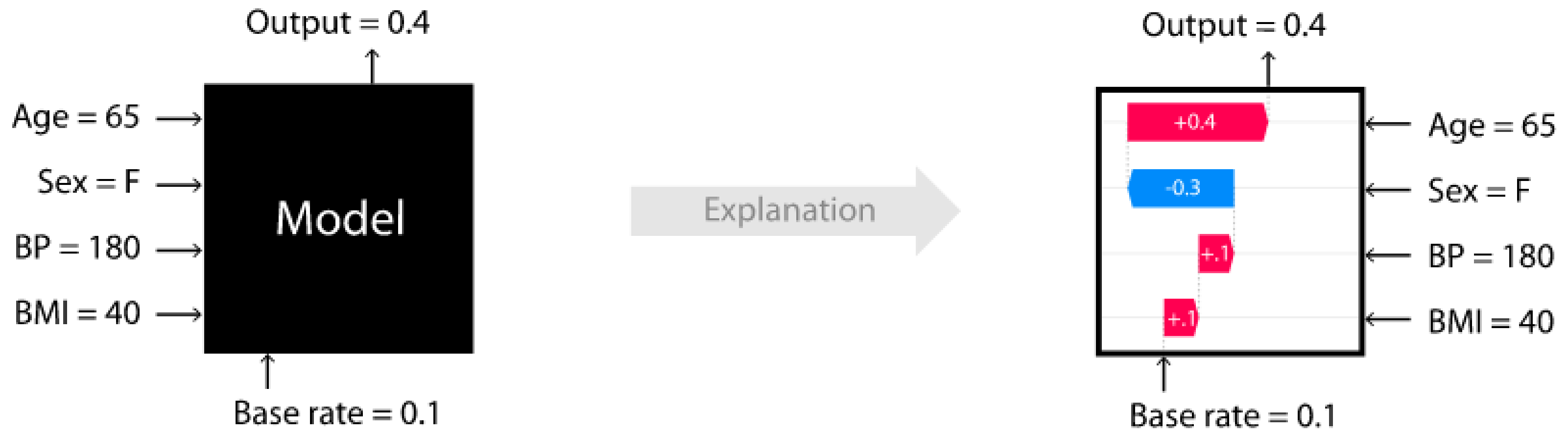
Shapley Values

- Оценим, насколько выигрыш обусловлен участием того или иного игрока
- (4+0+4+4+4+3) / 6 ~ 3,16 доля игрока A
- (7+7+10+3+9+10) / 6 ~ 7,66 доля игрока B
- (8+12+5+12+6+6) / 6 ~ 8,16 доля игрока C

https://clearcode.cc/blog/game-theory-attribution/

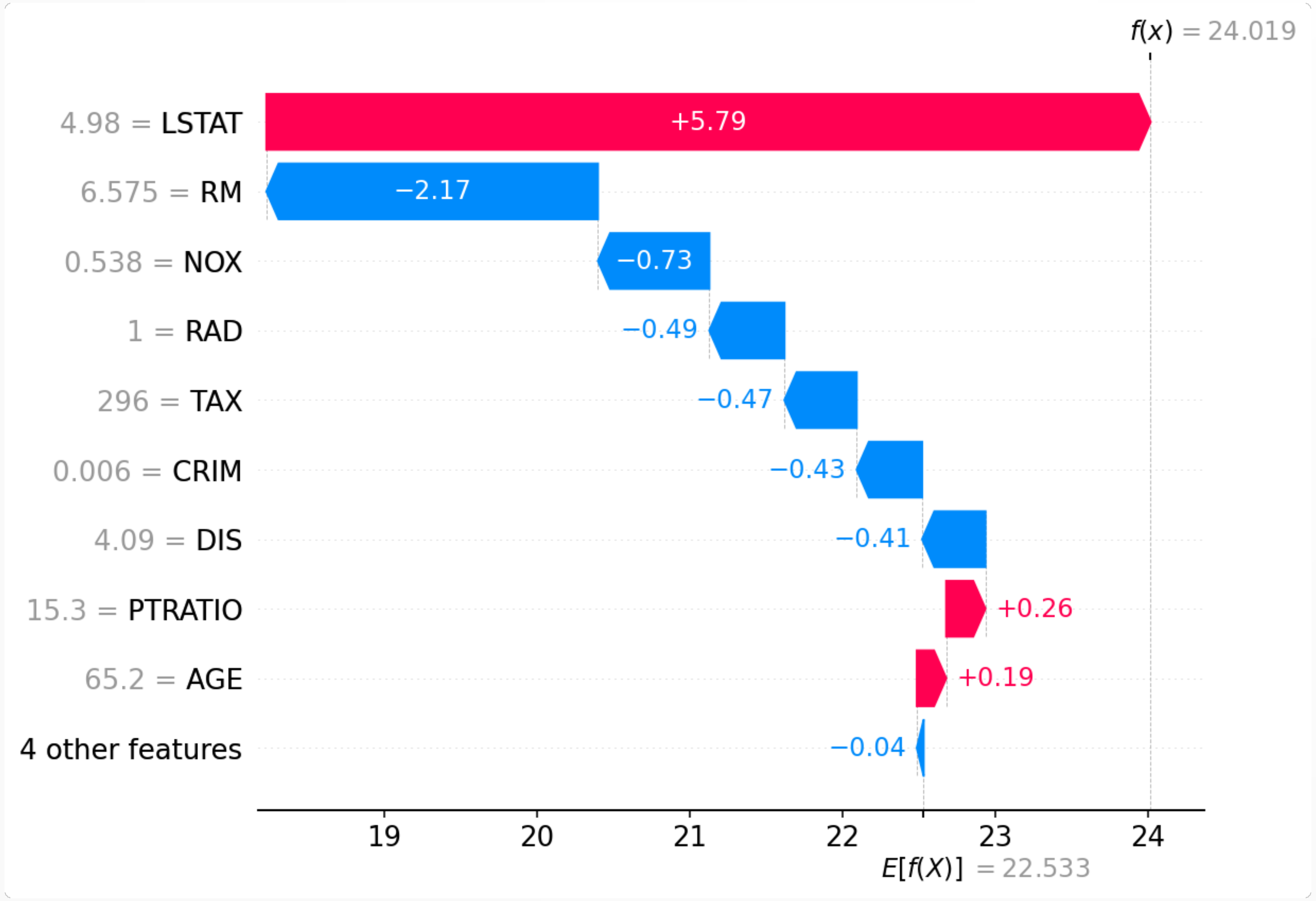
Признаки как игроки

SHAP


Output = 0.4
Age = 65
Sex = F
BP = 180
BMI = 40
Model
Base rate = 0.1
Explanation
Output = 0.4
+0.4
-0.3
+.1
+.1
Age = 65
Sex = F
BP = 180
BMI = 40
Base rate = 0.1

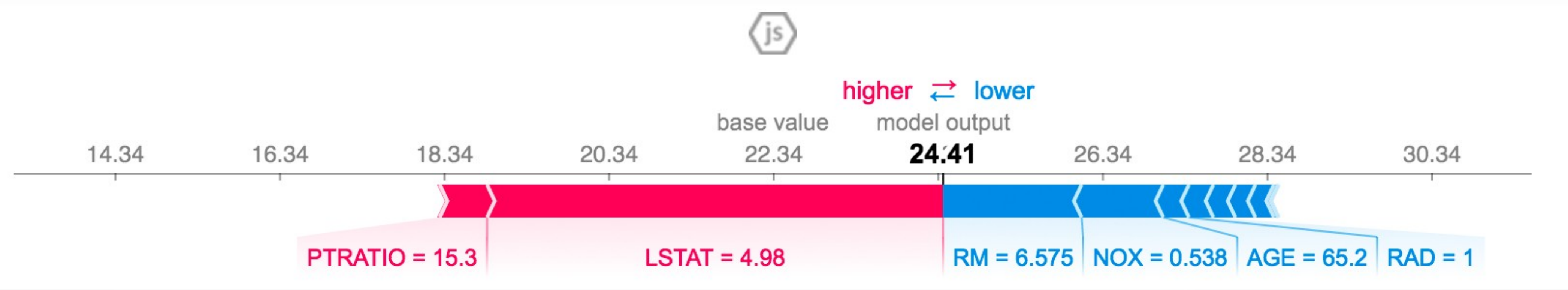
https://github.com/slundberg/shap

Водопад



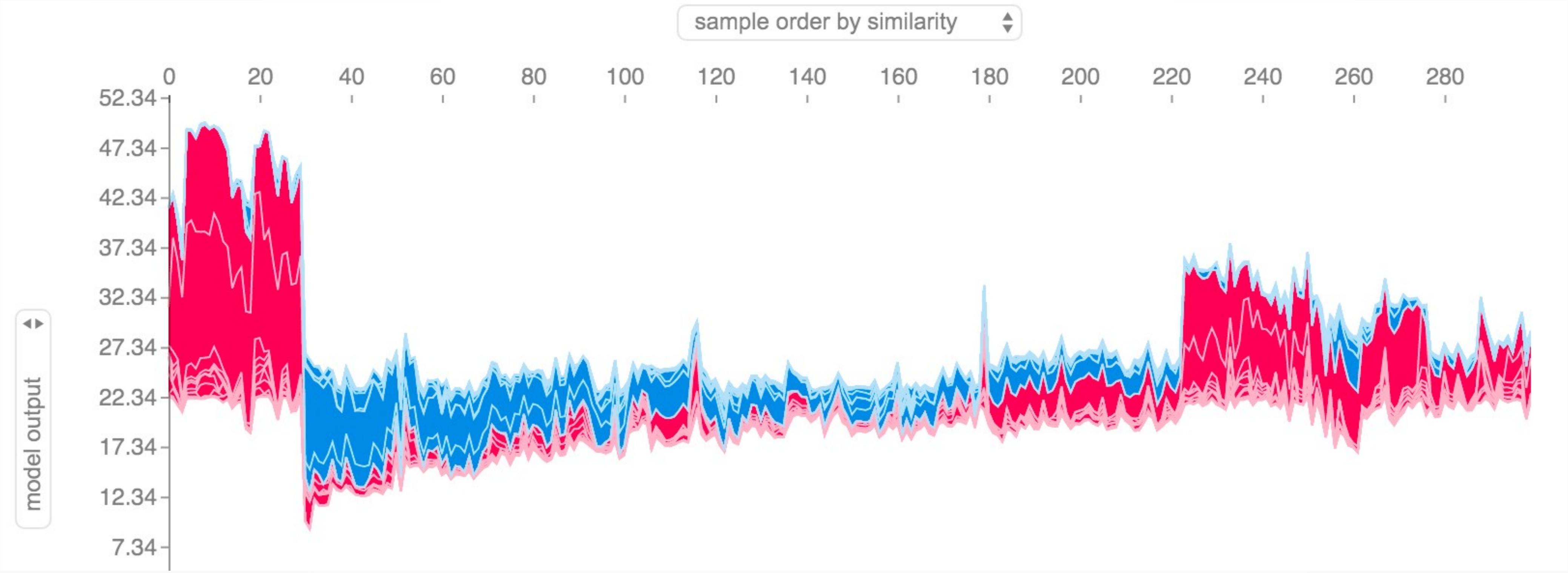
https://github.com/slundberg/shap

Force plot



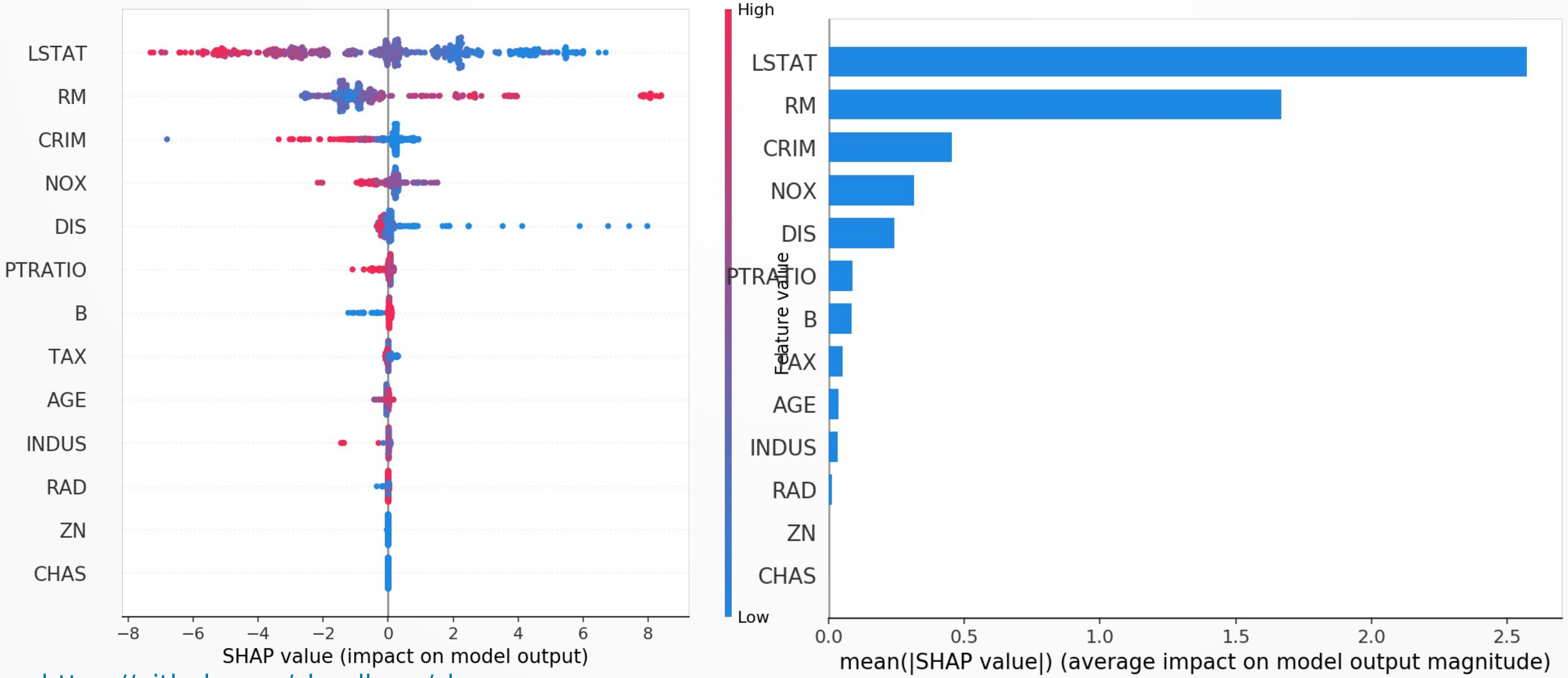
https://github.com/slundberg/shap

SHAP и структура данных



https://github.com/slundberg/shap

SHAP — глобальный



https://github.com/slundberg/shap

Учтите

https://arxiv.org/pdf/2002.11097.pdf

Problems with Shapley-value-based explanations as feature importance measures

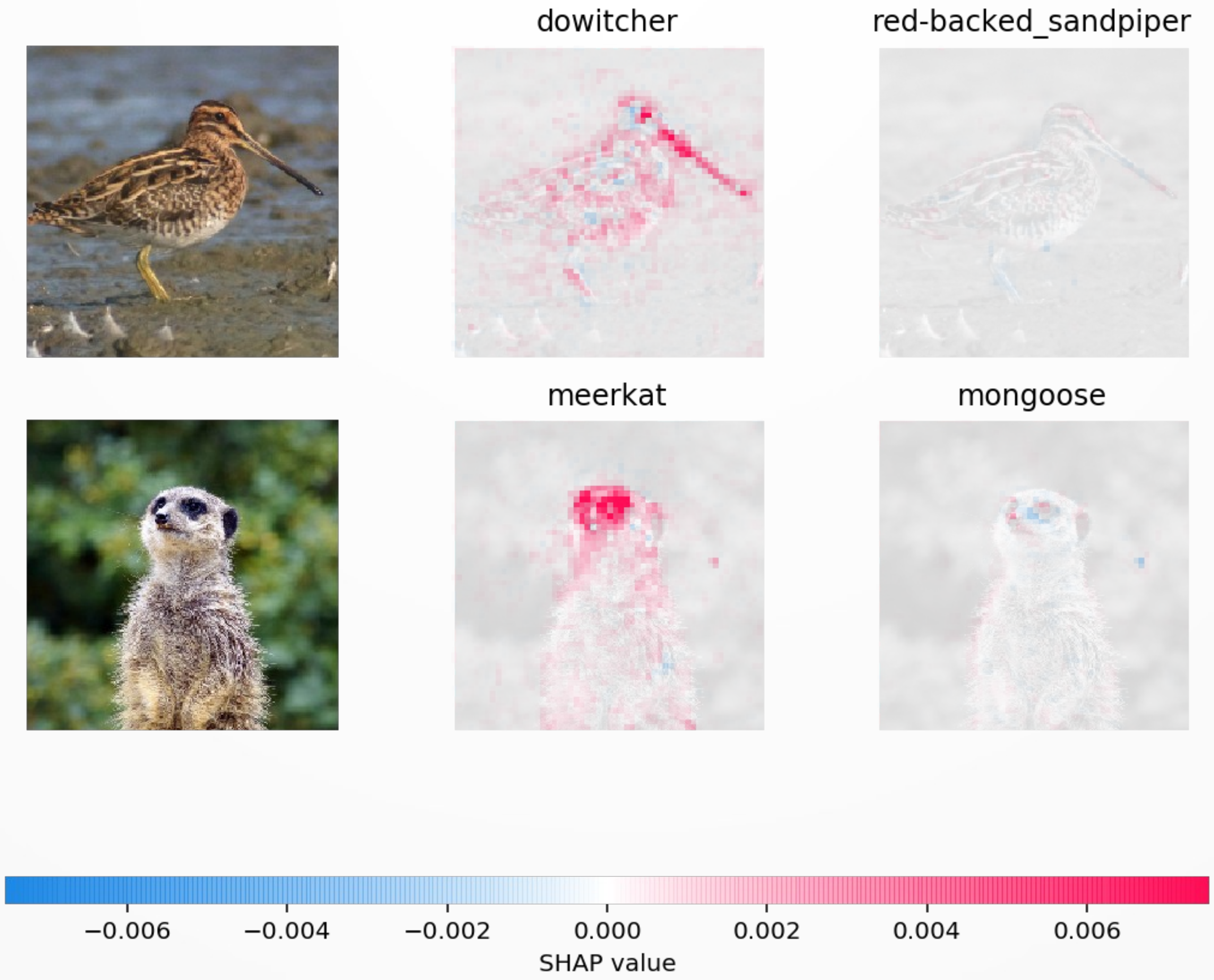
https://papers.nips.cc/paper/2020/file/0d770c496aa3da6d2c3f2bd19e7b9d6b-Paper.pdf

Asymmetric Shapley values

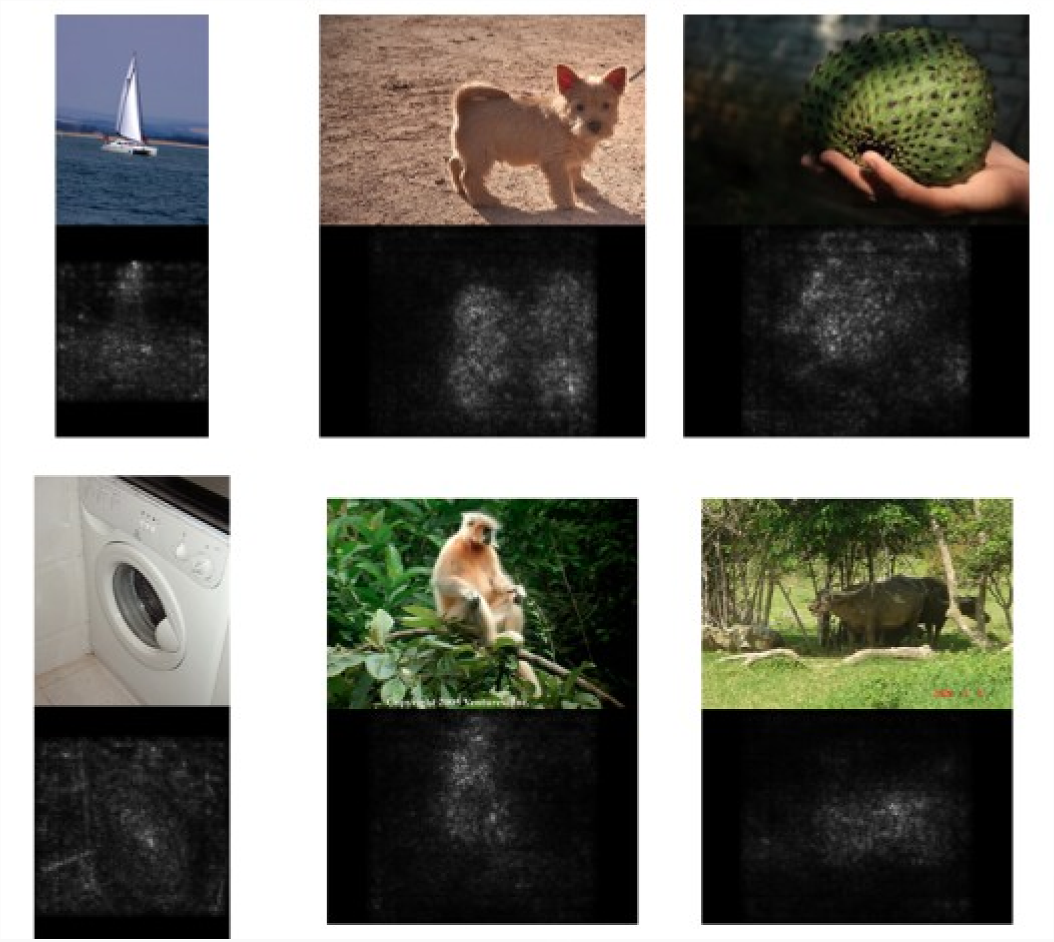
https://papers.nips.cc/paper/2020/file/32e54441e6382a7fbacbbbaf3c450059-Paper.pdf

Causal Shapley Values

SHAP — влияющие точки



Saliency map



- Подсветим точки градиентом
- Можно наложить маску на исходное изображение
- На что обратила внимание != Чем руководствовалась

https://arxiv.org/abs/1312.6034

http://www.scholarpedia.org/article/Saliency_map

Saliency map

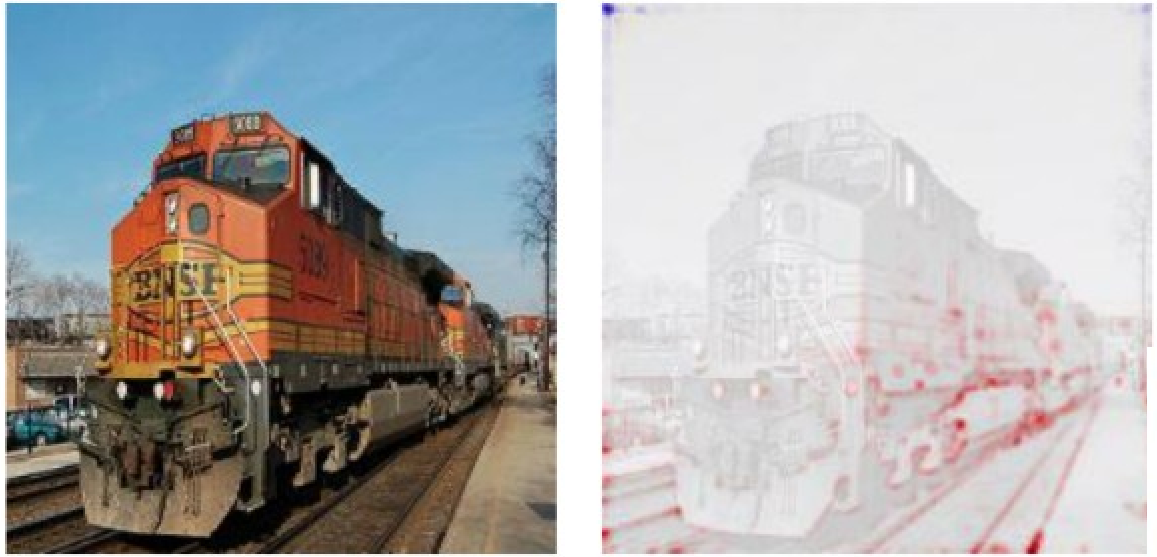


Figure 2: Example of a saliency map explanation of a True Positive (TP) image for the label *"train"*. It highlights the contours of the lines below the train. A possible interpretation is that the CNN has learned to recognise trains when rails are present.

by the CNN. Our results indicate that when saliency maps were available, participants answered correctly more frequently than when they were absent (60.7% vs. 55.1%, p = 0.045). However, the overall performance was generally low even with the presence of saliency maps.

Our data indicates that saliency maps influenced people to notice saliency-maps-features. However, it is unclear if such explanation draw them away from considering other attributes that are usually not highlighted by saliency maps.

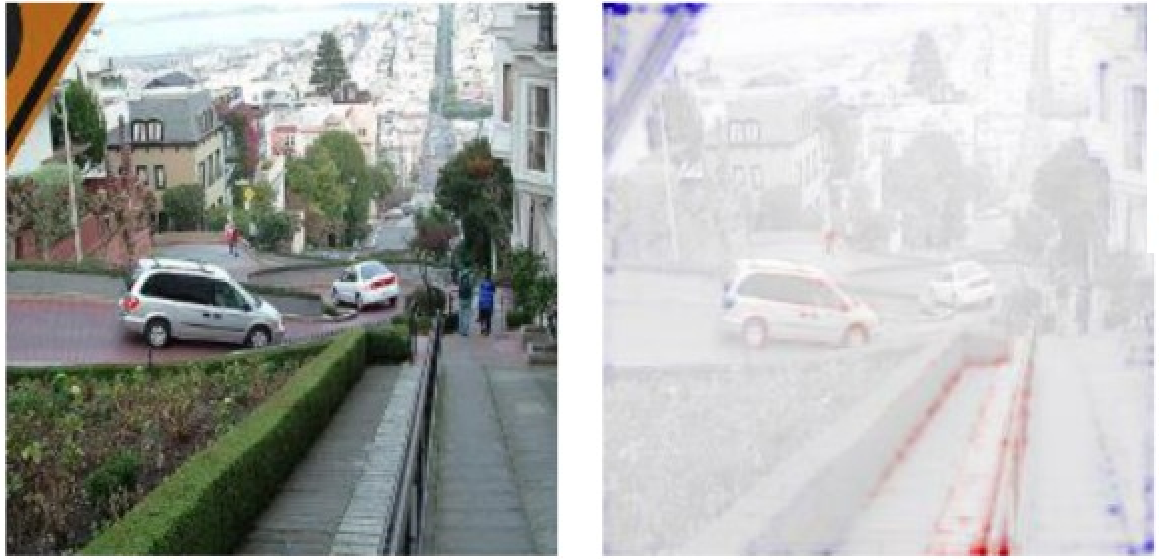
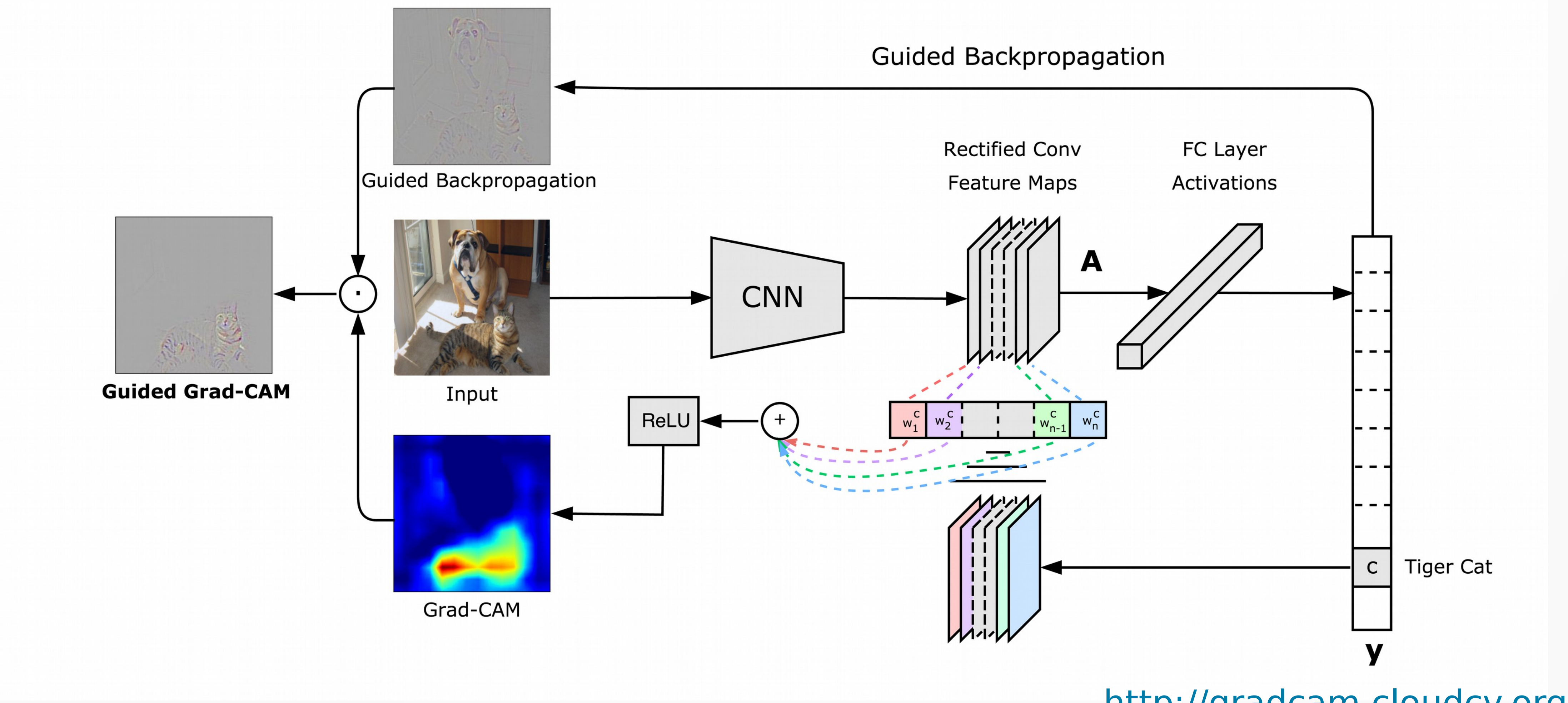


Figure 3: Example of a saliency map explanation of a False Positive (FP) image for the label *"train"*. A possible interpretation is that edges in the lower part appeared similar to rails, which could explain this error.

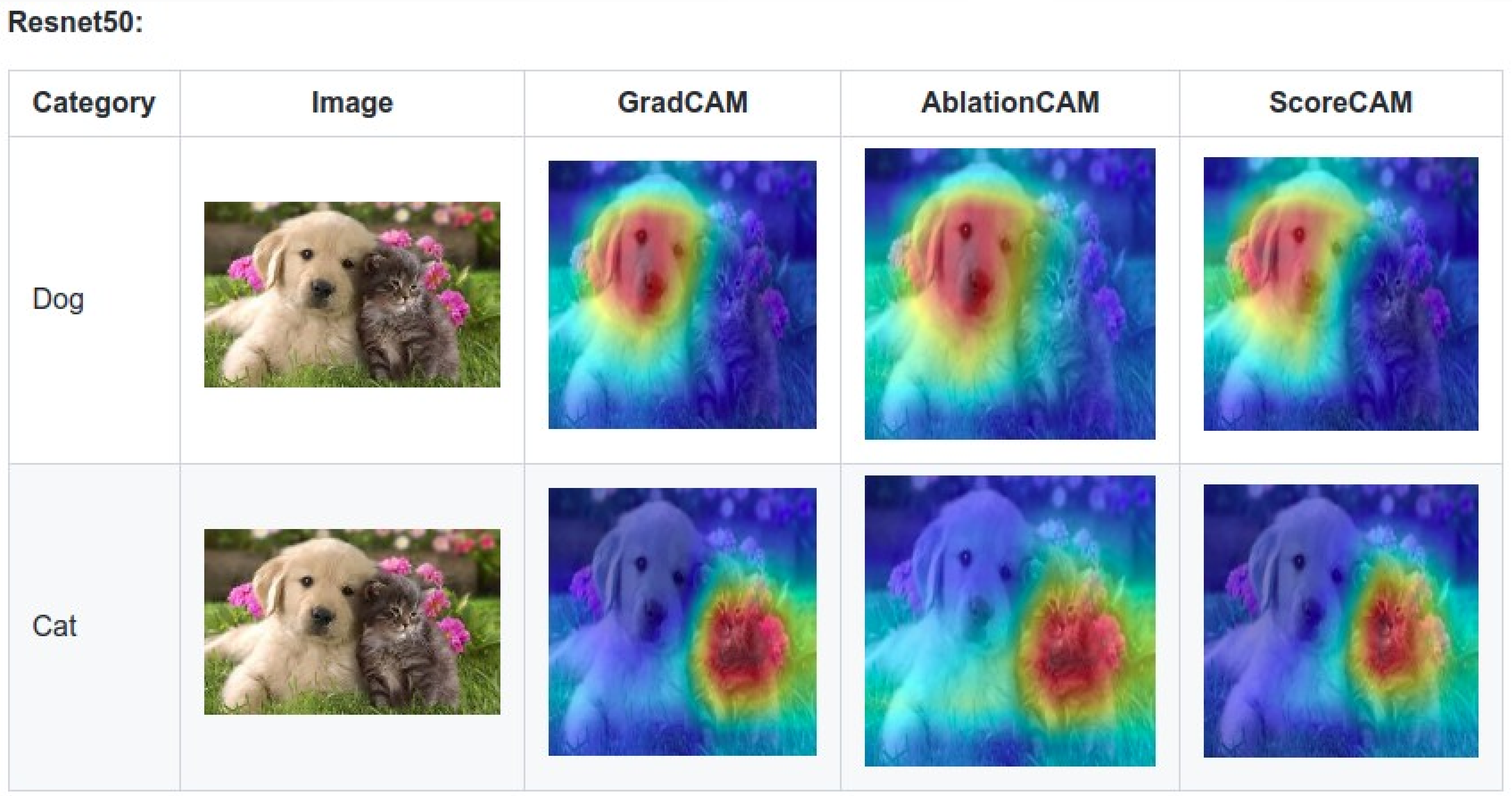
https://arxiv.org/pdf/2002.00772.pdf

Grad-CAM


Guided Backpropagation
Guided Backpropagation
Rectified Conv
Feature Maps
FC Layer
Activations
A
CNN
Input
Guided Grad-CAM
ReLU
+
w^c_1
w^c_2
w^c_{n-1}
w^c_n
Grad-CAM
c
Tiger Cat
y

http://gradcam.cloudcv.org/

Grad-CAM



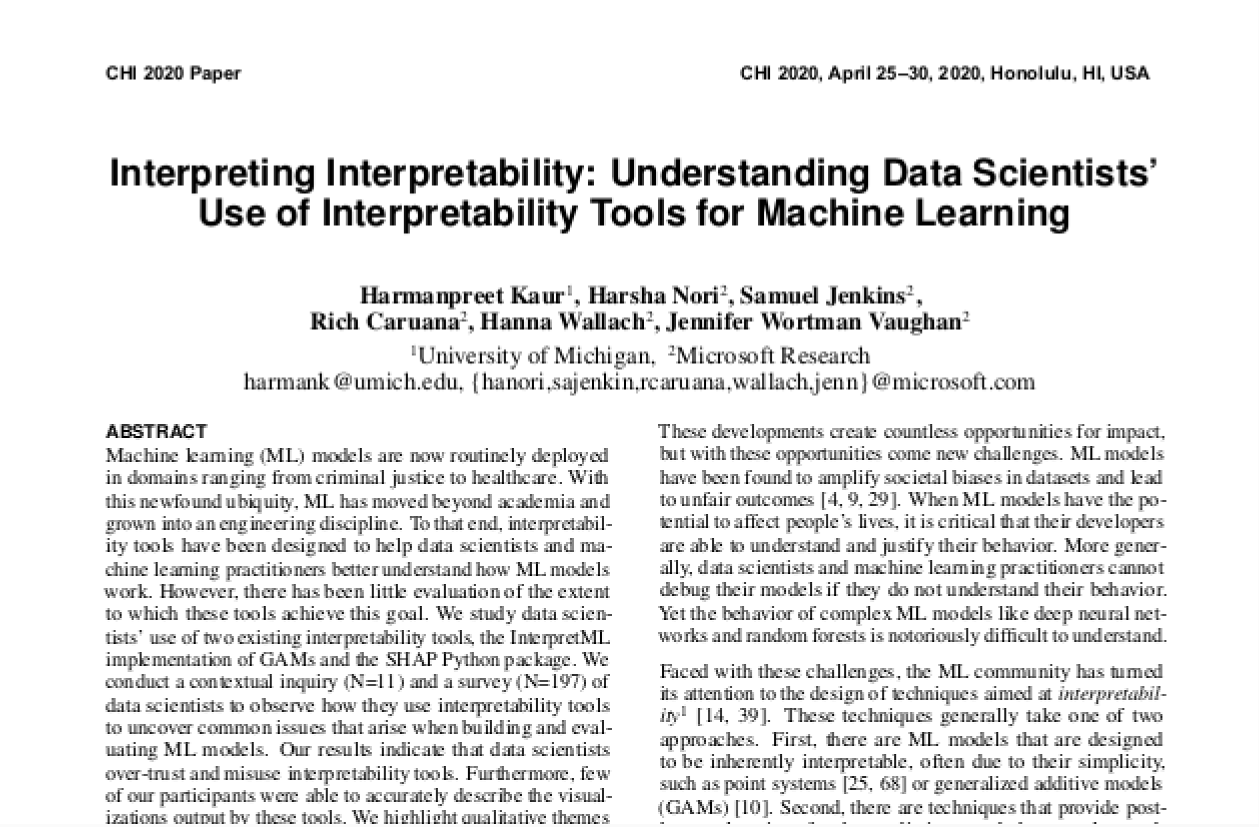
Resnet50:

Category	Image	GradCAM	AblationCAM	ScoreCAM
Dog				
Cat				

https://github.com/jacobgil/pytorch-grad-cam

Лишь бы график красивый

- У нас был датасет
- Пачка графиков
- 11 датасатанистов
- SHAP
- GAM
- …



Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning

Harmanpreet Kaur[1], **Harsha Nori**[2], **Samuel Jenkins**[2], **Rich Caruana**[2], **Hanna Wallach**[2], **Jennifer Wortman Vaughan**[2]
[1]University of Michigan, [2]Microsoft Research
harmank@umich.edu, {hanori,sajenkin,rcaruana,wallach,jenn}@microsoft.com

ABSTRACT

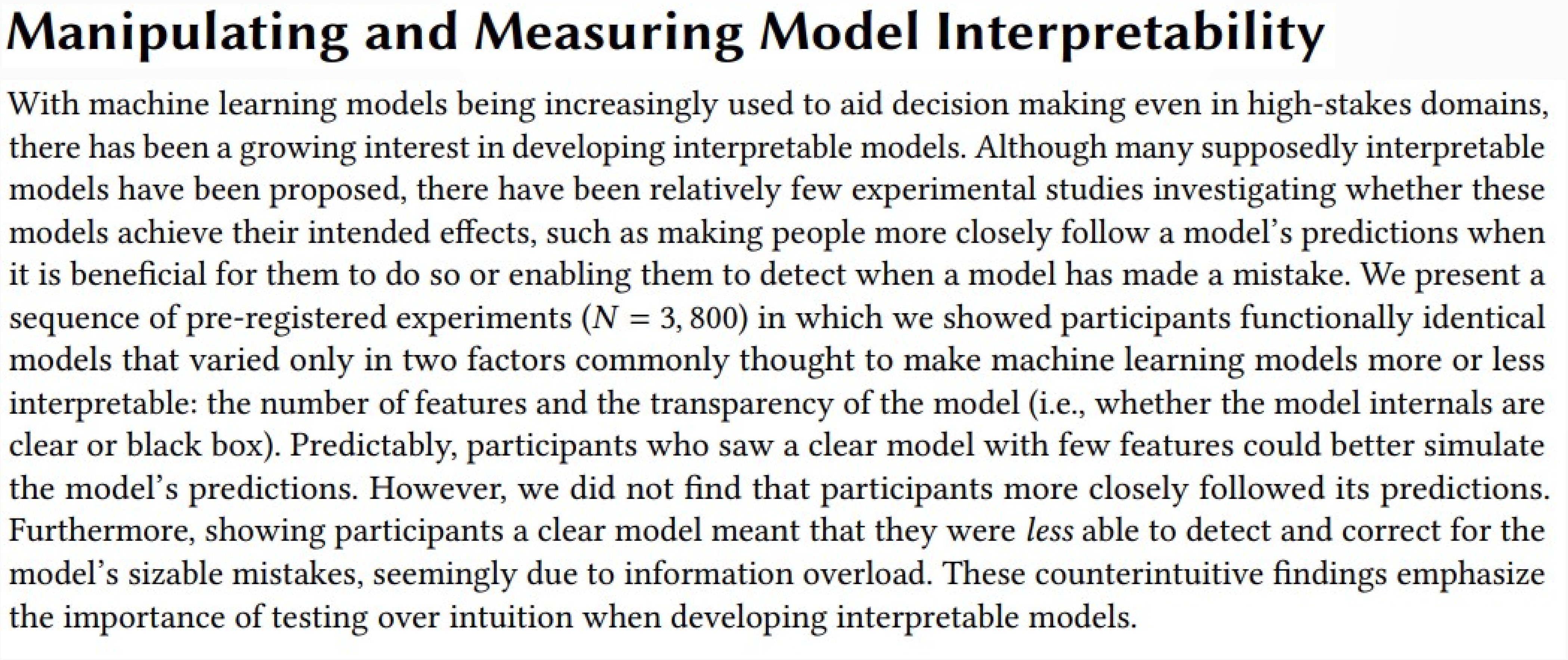
Machine learning (ML) models are now routinely deployed in domains ranging from criminal justice to healthcare. With this newfound ubiquity, ML has moved beyond academia and grown into an engineering discipline. To that end, interpretability tools have been designed to help data scientists and machine learning practitioners better understand how ML models work. However, there has been little evaluation of the extent to which these tools achieve this goal. We study data scientists' use of two existing interpretability tools, the InterpretML implementation of GAMs and the SHAP Python package. We conduct a contextual inquiry (N=11) and a survey (N=197) of data scientists to observe how they use interpretability tools to uncover common issues that arise when building and evaluating ML models. Our results indicate that data scientists over-trust and misuse interpretability tools. Furthermore, few of our participants were able to accurately describe the visualizations output by these tools. We highlight qualitative themes

These developments create countless opportunities for impact, but with these opportunities come new challenges. ML models have been found to amplify societal biases in datasets and lead to unfair outcomes [4, 9, 29]. When ML models have the potential to affect people's lives, it is critical that their developers are able to understand and justify their behavior. More generally, data scientists and machine learning practitioners cannot debug their models if they do not understand their behavior. Yet the behavior of complex ML models like deep neural networks and random forests is notoriously difficult to understand.

Faced with these challenges, the ML community has turned its attention to the design of techniques aimed at *interpretability*[1] [14, 39]. These techniques generally take one of two approaches. First, there are ML models that are designed to be inherently interpretable, often due to their simplicity, such as point systems [25, 68] or generalized additive models (GAMs) [10]. Second, there are techniques that provide post-

http://www.jennwv.com/papers/interp-ds.pdf

Прозрачное хуже



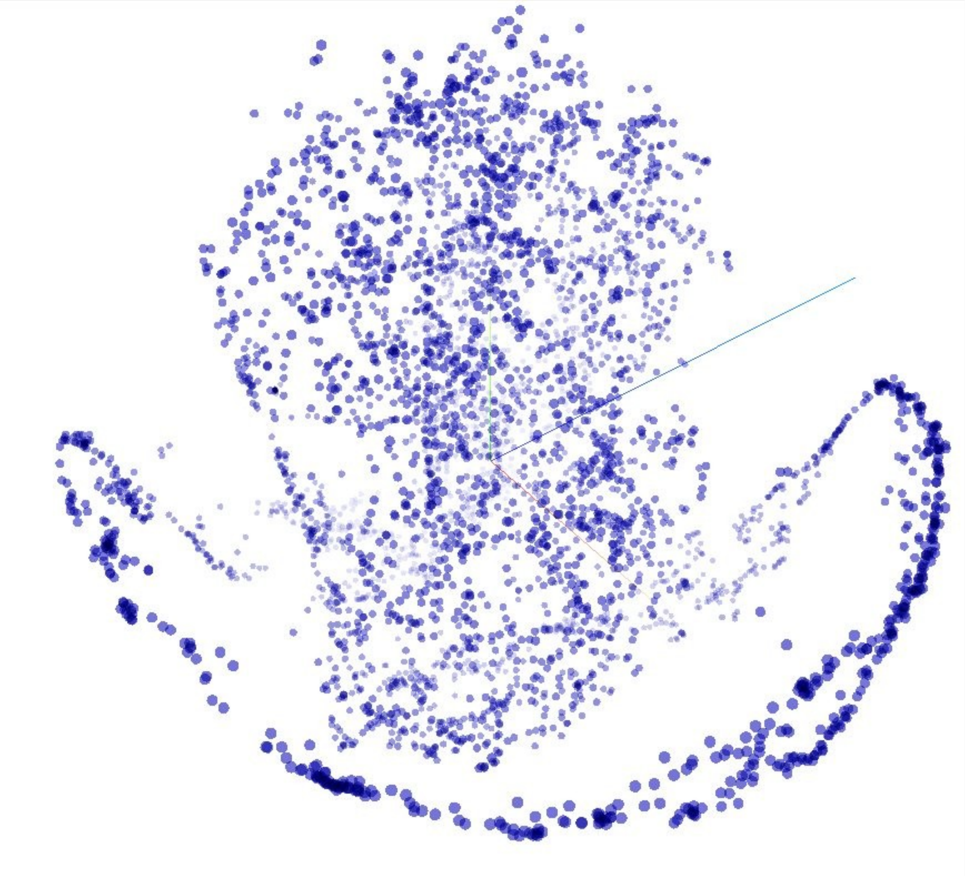
Manipulating and Measuring Model Interpretability

With machine learning models being increasingly used to aid decision making even in high-stakes domains, there has been a growing interest in developing interpretable models. Although many supposedly interpretable models have been proposed, there have been relatively few experimental studies investigating whether these models achieve their intended effects, such as making people more closely follow a model's predictions when it is beneficial for them to do so or enabling them to detect when a model has made a mistake. We present a sequence of pre-registered experiments ($N = 3,800$) in which we showed participants functionally identical models that varied only in two factors commonly thought to make machine learning models more or less interpretable: the number of features and the transparency of the model (i.e., whether the model internals are clear or black box). Predictably, participants who saw a clear model with few features could better simulate the model's predictions. However, we did not find that participants more closely followed its predictions. Furthermore, showing participants a clear model meant that they were *less* able to detect and correct for the model's sizable mistakes, seemingly due to information overload. These counterintuitive findings emphasize the importance of testing over intuition when developing interpretable models.

https://arxiv.org/pdf/1802.07810.pdf

Еще не все

- Интерпретация эмбеддингов
  - Projector
  - Граф близости эталонов
- Графовые модели
- Трансформеры и внимание
- Топологические ML-модели
- Causal Inference
- Reinforcement Learning



Что делать

- Не пропускать этап EDA
- Разрабатывать инструменты
  - Есть пара идей для дипломной работы
- Учиться интерпретировать модели
  - Курс от ODS.AI, стартует в октябре
- Думать, что делаешь

Вопросы

Слайды тут

dkolodezev

promsoft

dkolodezev

d_key

dmitry_kolodezev

https://kolodezev.ru/download/interpretable20210923.pdf