

Интерпретация моделей машинного обучения

Лекция 10
Дмитрий Колодезев
НГУ 2020, весна

План

- Проблемы черных ящиков
- Понятные модели
- Непонятные модели
- LIME — локально линейно
- SHAP — кто виноват
- Проблемы нейронок
- GRAD-CAM

Проблемы черных ящиков

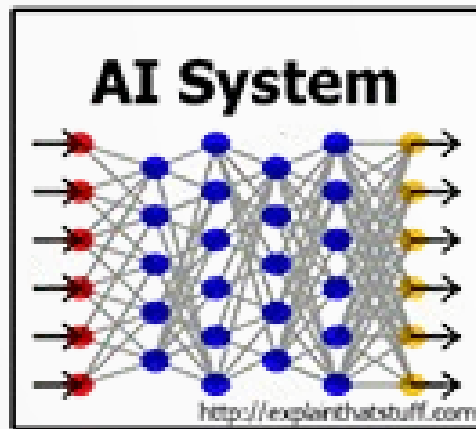
- Вот лампочка. Вот провод. Вот выключатель. Понятно?
- Все понятно, только как керосин по проводам в лампочку попадает?



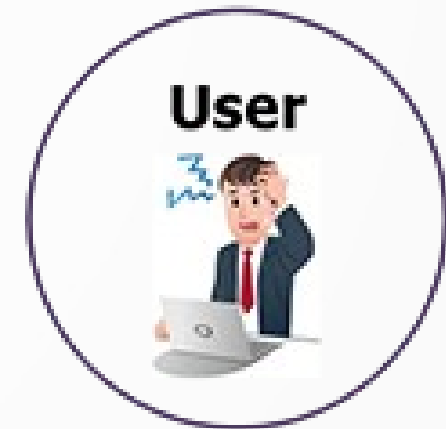
Зачем

- Чтобы отлаживать
- Чтобы доверять
- Чтобы предсказывать поведение
- Чтобы использовать знания
- Чтобы разбирать полеты

DARPA, eXplainable AI



- 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



- 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?

Когда не нужно

- Влияние модели мало
- Проблема хорошо разработана
- Класс моделей широко применяется
 - линейные модели
- Хотим скрыть алгоритм
 - скоринг
 - ранжирование
 - оценка качества

Интерпретируемые модели

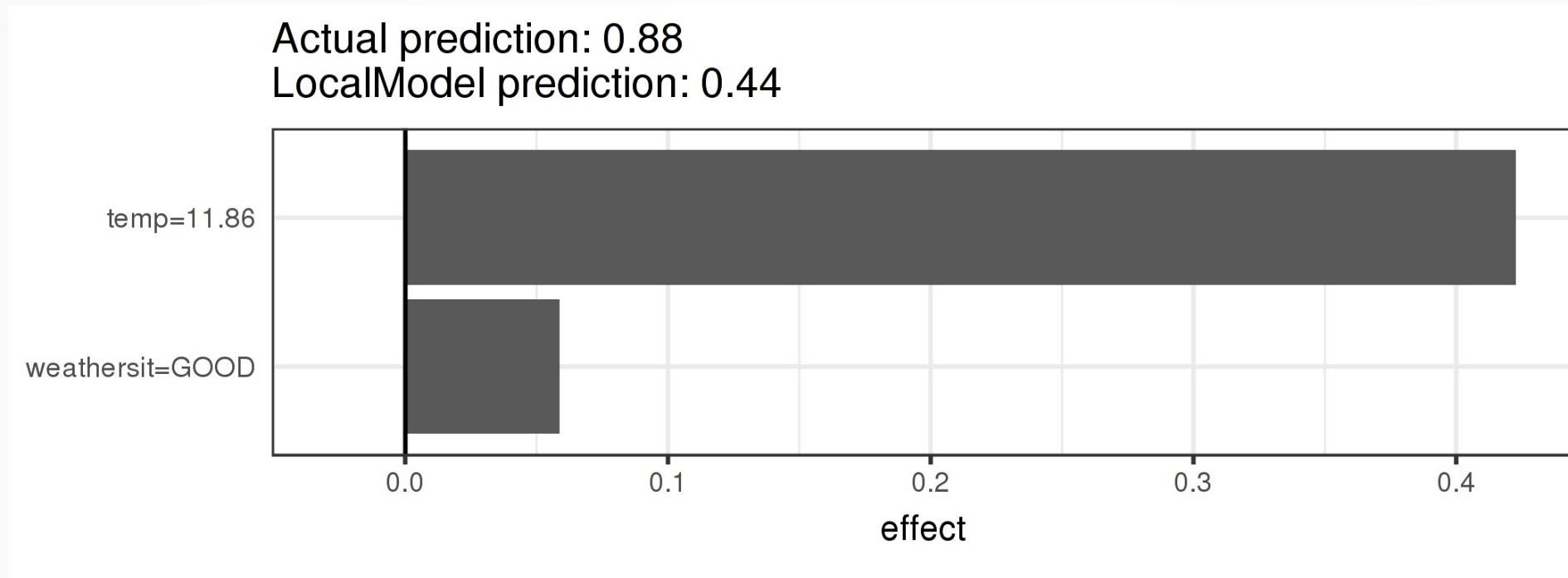
- Линейные модели
 - Сравнивать эффекты, не веса
 - Иллюзия плотности
 - One Hot Encoding и т.д.
 - Взаимодействие признаков
- Деревья (неглубокие) и списки правил
- kNN и на примерах
 - Когда точки интерпретируемы

Варианты

- Только интерпретируемые модели, только LR
- Интерпретируемые поверх эмбедингов
- Изучаем веса, активации
- Внутренняя «важность»
- Суррогатные модели — например, LIME
- SHAP, или научный подход к делёжке
- Свои подходы для нейронных сетей

LIME — локальные суррогаты

- Локальная сурогатная модель
- Интерпретируемая селективная модель
- Может обучаться на других признаках



LIME — ТЕКСТ

- For Christmas Song visit my channel! ;)

case	label_prob	feature	feature_weight
1	0.1701170	good	0.000000
1	0.1701170	PSY	0.000000
1	0.1701170	a	0.000000
2	0.9939024	channel!	6.180747
2	0.9939024	Song	0.000000
2	0.9939024	Christmas	0.000000

LIME — суперпиксели



(a) Original Image



(b) Explaining *Electric guitar*



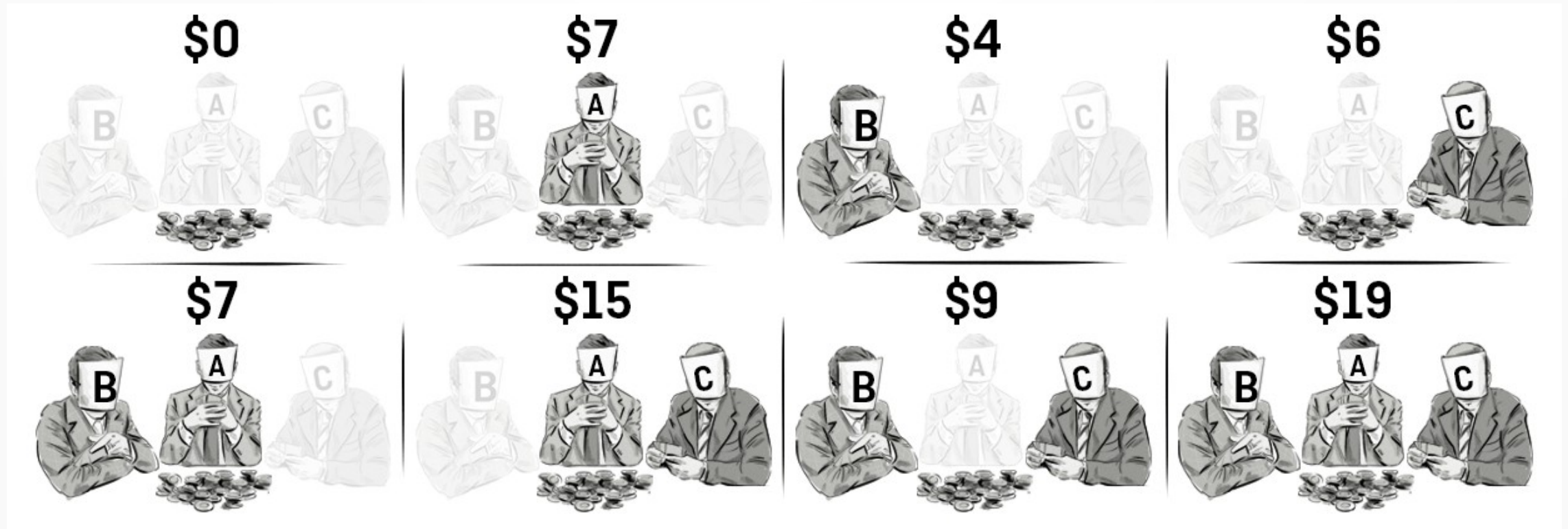
(c) Explaining *Acoustic guitar*



(d) Explaining *Labrador*

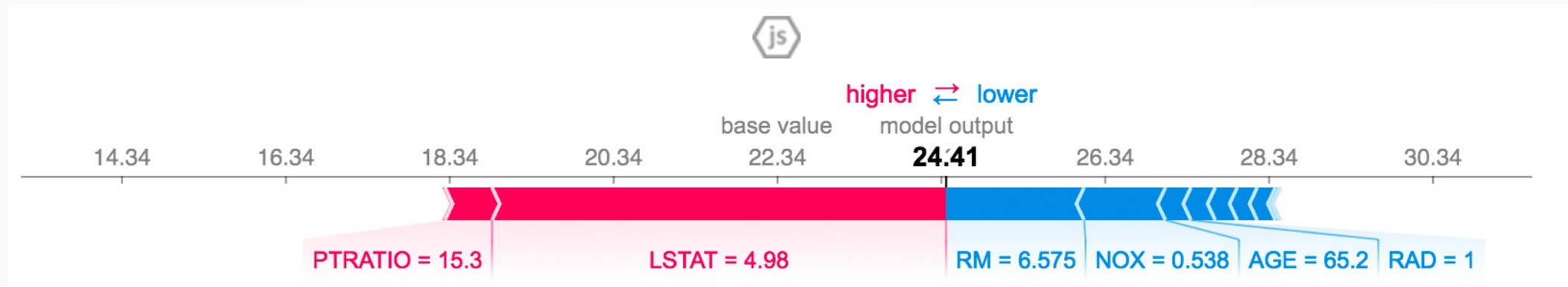
It's not a Lab in the picture, it's a Golden Retriever (a light colored one) — A.F.

Shapley Values

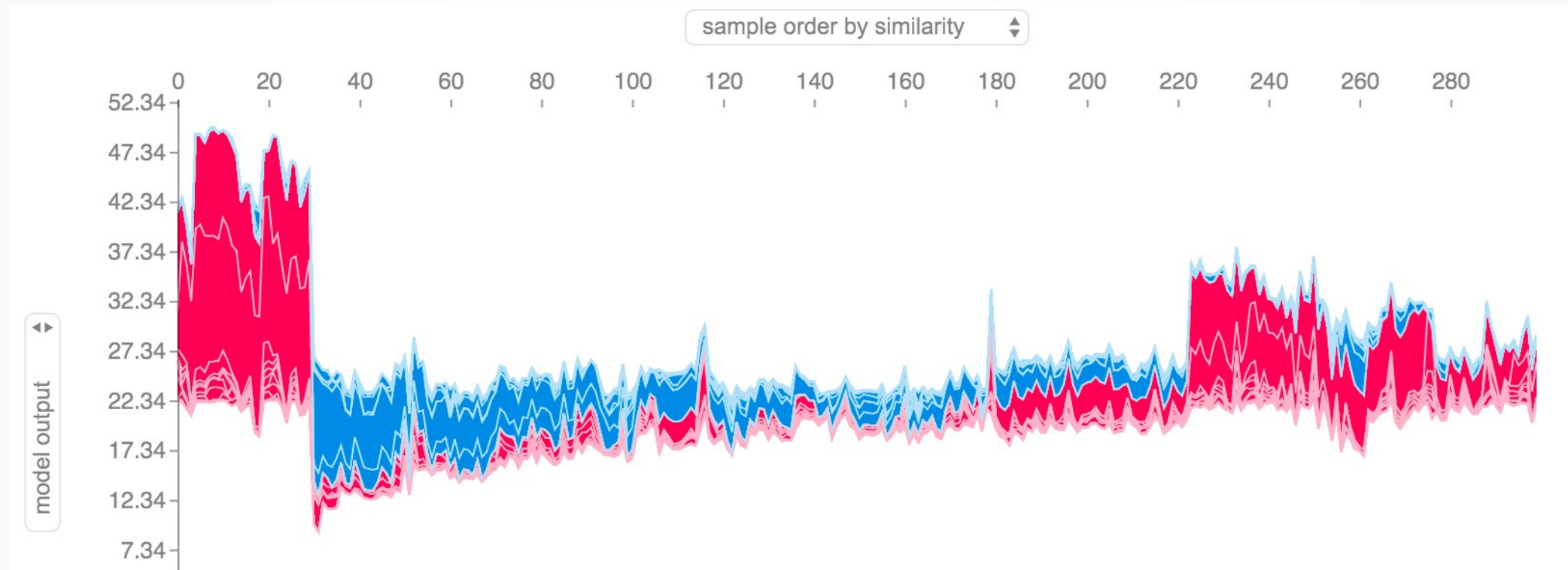


<https://clearcode.cc/blog/game-theory-attribution/>

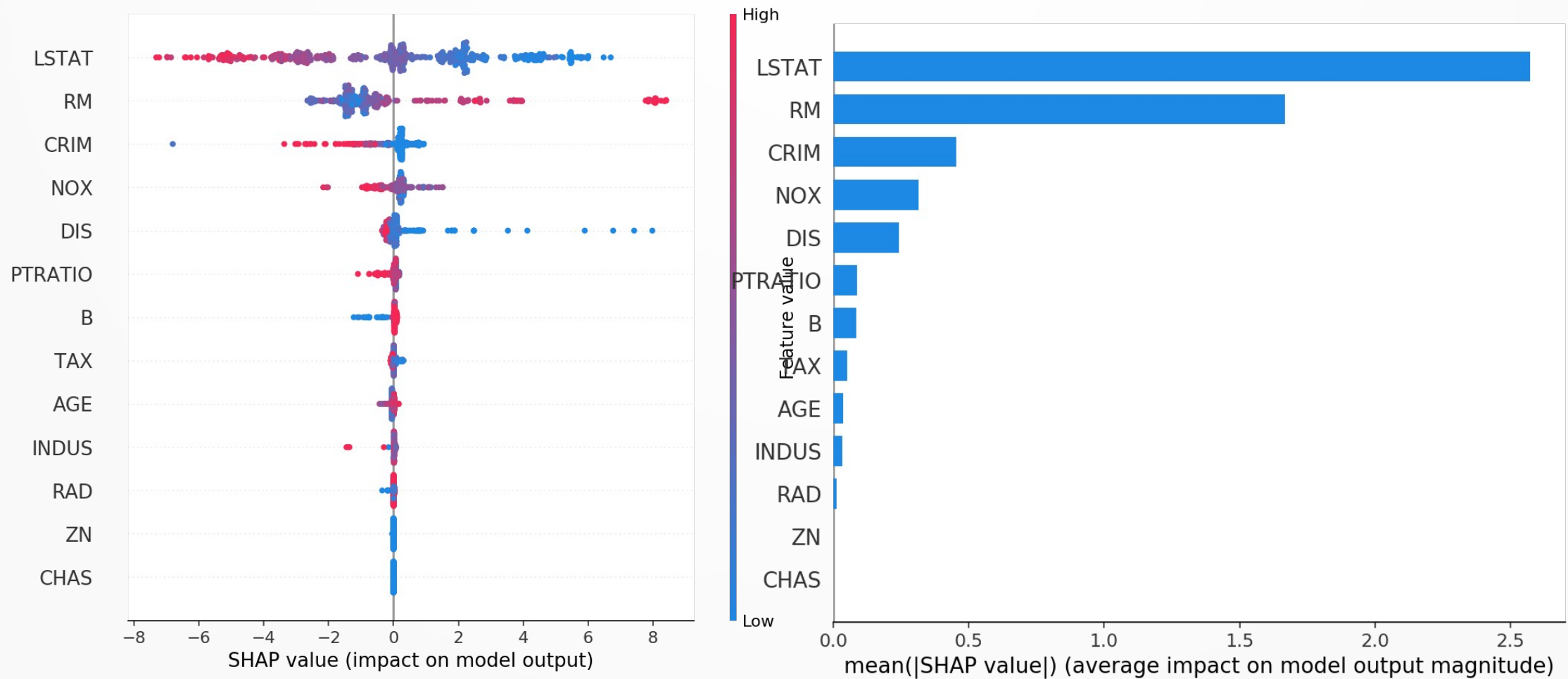
Shapley Values & SHAP



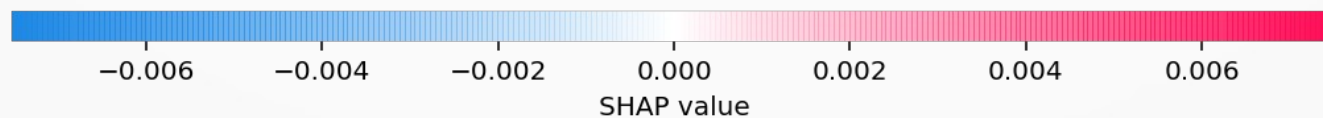
SHAP и структура данных



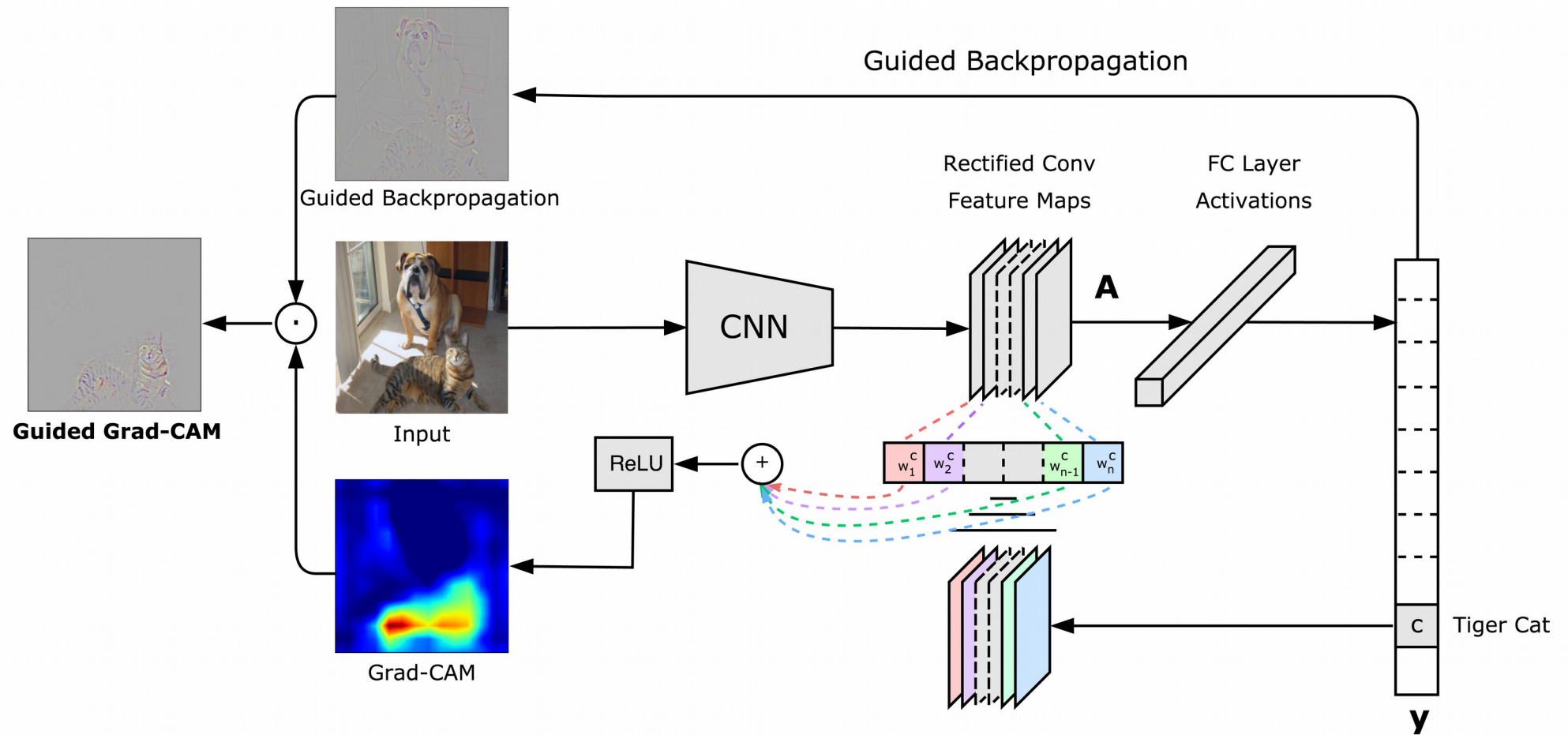
SHAP — глобальный



SHAP — ВЛИЯЮЩИЕ ТОЧКИ

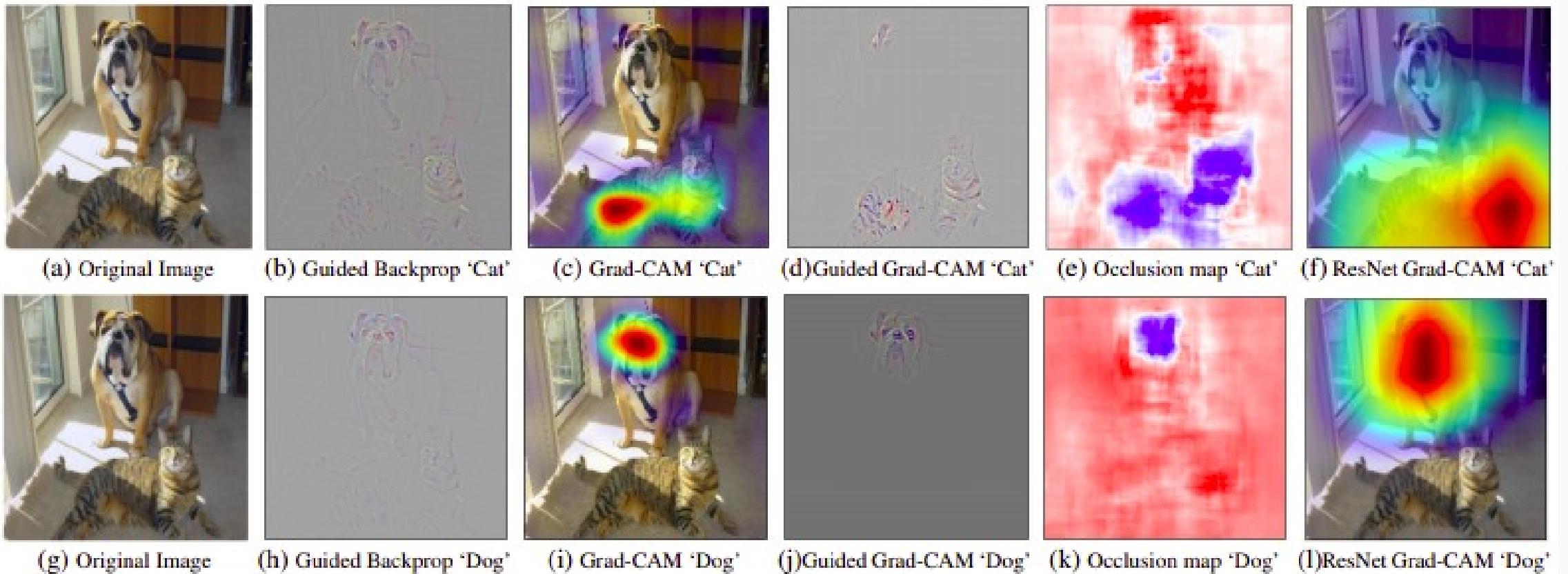


Grad-CAM



<http://gradcam.cloudcv.org/>

Grad-CAM



<https://arxiv.org/abs/1610.02391>

Почитать

- Дьяконов, Интерпретации чёрных-ящиков
- Becker, Machine Learning Explainability
- Molnar, Interpretable Machine Learning
- <https://github.com/marcotcr/lime>
- <https://github.com/slundberg/shap>
- Grad-CAM

Вопросы

Слайды тут



dkolodezev



promsoft



dkolodezev



d_key



dmitry_kolodezev

<https://kolodezev.ru/download/slides-interpretation-v5.pdf>