Как думают роботы. Интерпретация ML-моделей

Дмитрий Колодезев
ООО Промсофт, Новосибирск
28.05.2020 Сборка_

DK

- Первая ML-модель в 1986-м на Fortran 4, keep trying
- В Промсофте разрабатываем и внедряем ML-модели
- Преподаю, консультирую, In ODS We Trust

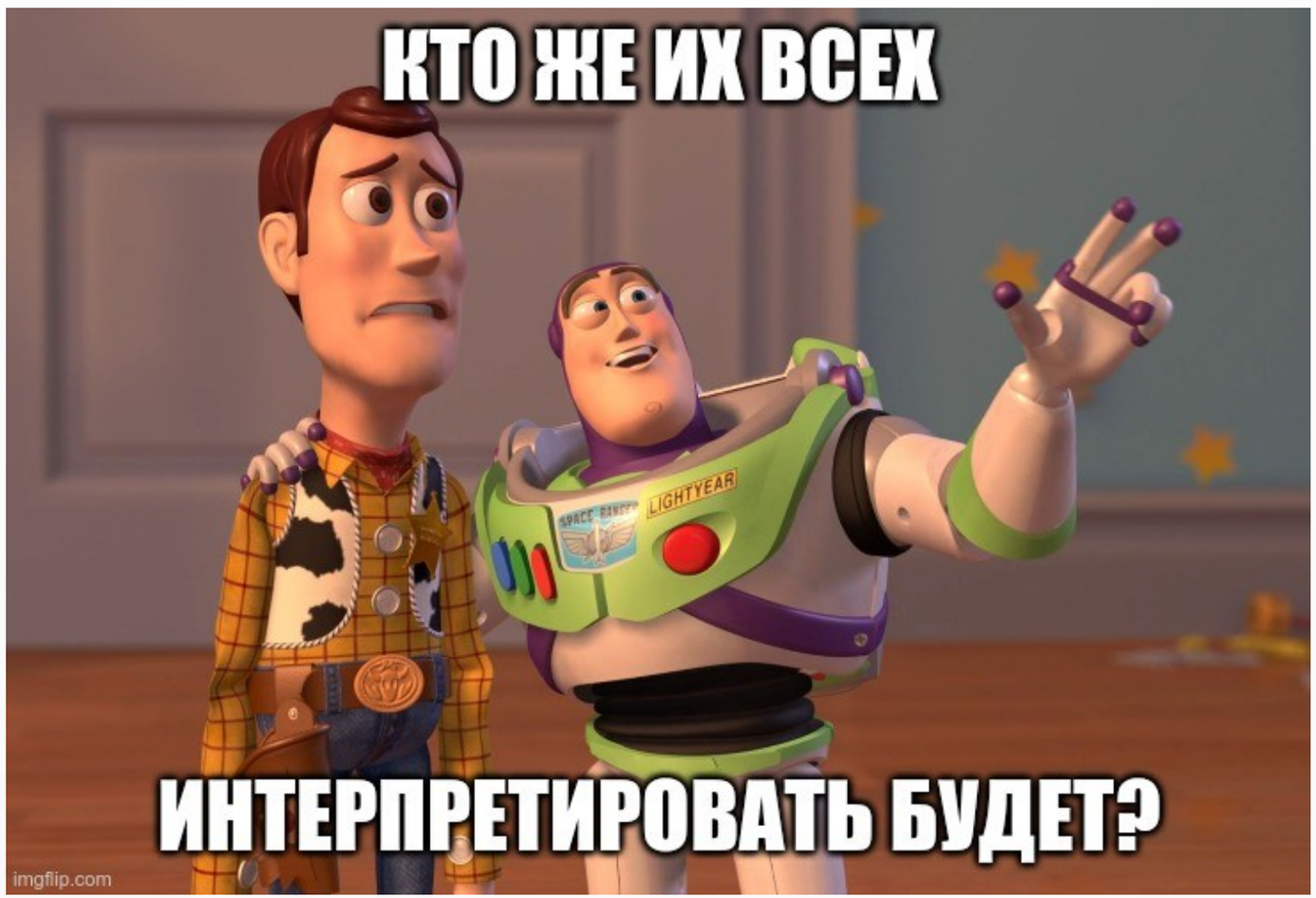
План

- О чем это все
- Кому и зачем это нужно
- В чем сложность
- Как сделать хорошо
- Как делают это с табличками
- Как делают это с картинками
- Что почитать и вопросы

Вещи стали умнее

- Алиса, Маруся и их заграничные подружки
- Умные розетки
- Тупые бухгалтерские программы
- Самодвижущиеся машины
- Автоматические штрафы за превышение
- Кредитный, кадровый и судебный скоринг
- Рекомендательные системы просто везде

Алоритмы заменяют людей

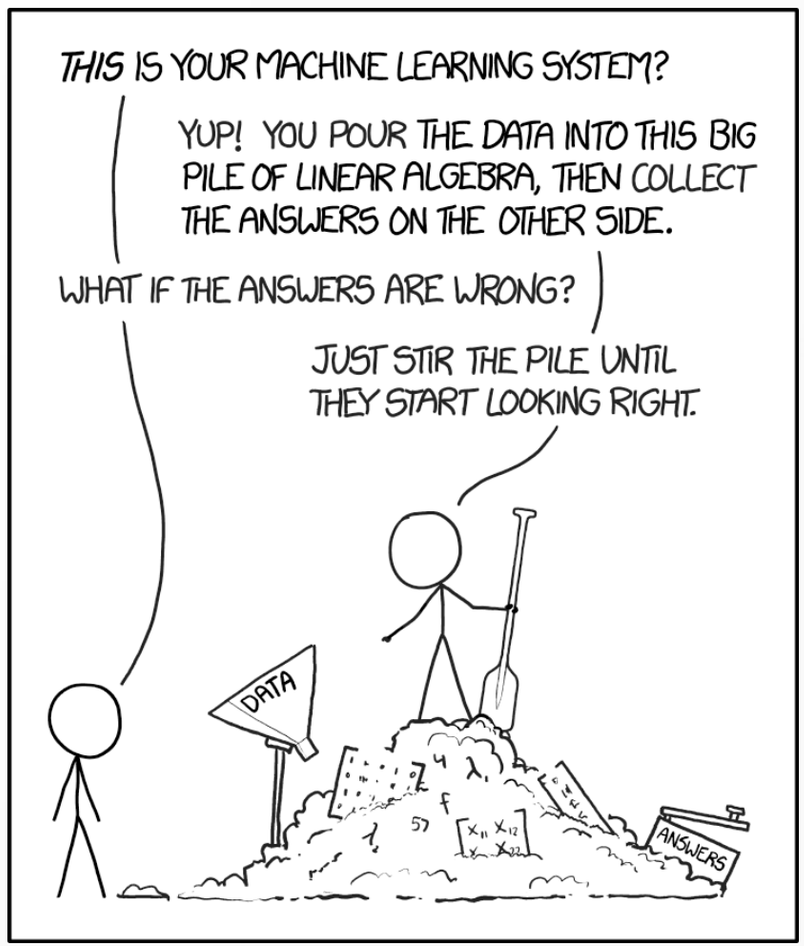


Как было в древности

- Люди собрали данные
- Выявили зависимости
- Написали программу
- Где-то ошиблись, что-то протестировали
- Почти не глючит
- Мы понимаем, как оно работает!
- Особенно пока программа маленькая

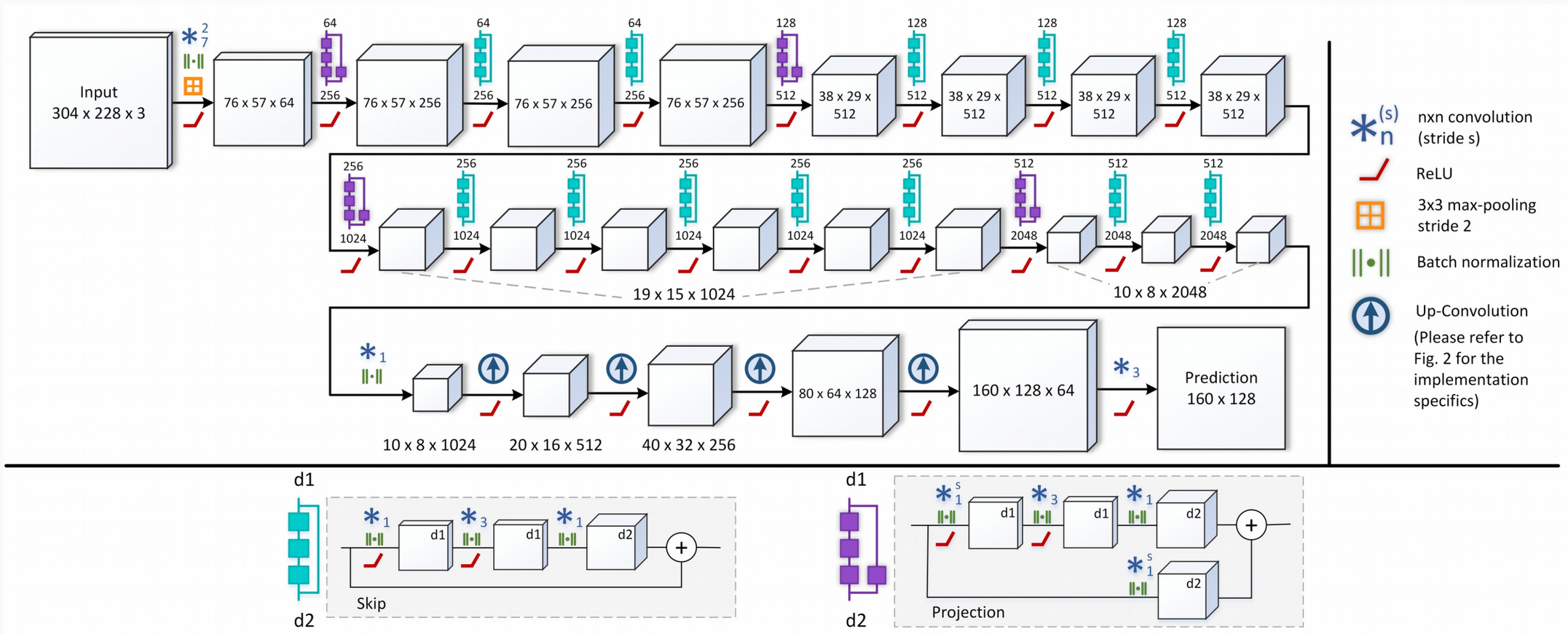
Как сейчас

- Собрали примеры
- Как-то разметили
- Перемешали лопатой
- Автоматически сгенерировали программу
- Она что-то выдает
- Ура!



Буквально — первые рекуррентные нейронные сети так и работали

В принципе все понятно



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

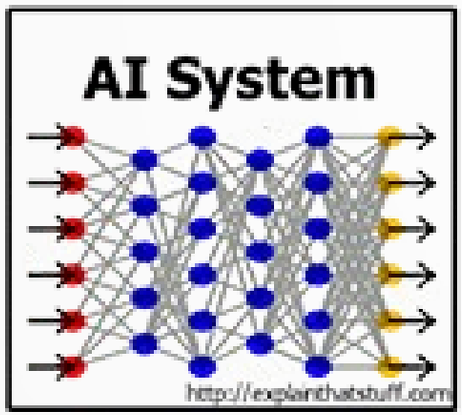
Программирование 2.0

- Andrej Karpathy, Software 2.0
- Хорошо:
  - Можем быстро делать сложные программы
- Плохо:
  - Никто толком не понимает, как они работают
- Впрочем:
  - Мы и раньше не понимали, но раньше для этого нужно было несколько лет работы.

Зачем это все понимать?

- Оно жульничает
- Оно учится плохому
- Оно странно работает
- Оно непредсказуемо
- Оно что-то знает
- Его бы улучшить
- А чего оно вообще...

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

DoD and non-DoD Applications

Transportation

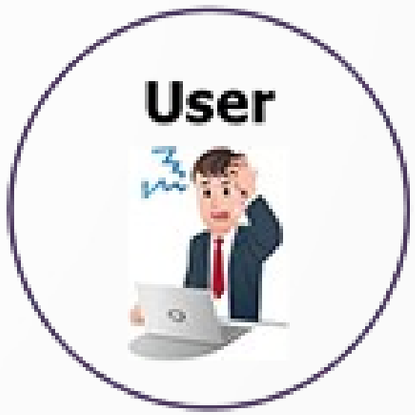
Security

Medicine

Finance

Legal

Military



- 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)

Когда не нужно

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

Что вообще бывает

- kNN
- Линейные модели
- Деревья
- Ансамбли моделей
- Нейронные сети
- ...

Хорошее объяснение

- Локальное — почему ему не дали кредит
- Глобальное — кому вообще дают кредиты
- Селективное — только важные признаки
- Непротиворечивое — нет контрпримеров
- Информативное — можно что-то сделать
- Понятное для пользователя

Варианты

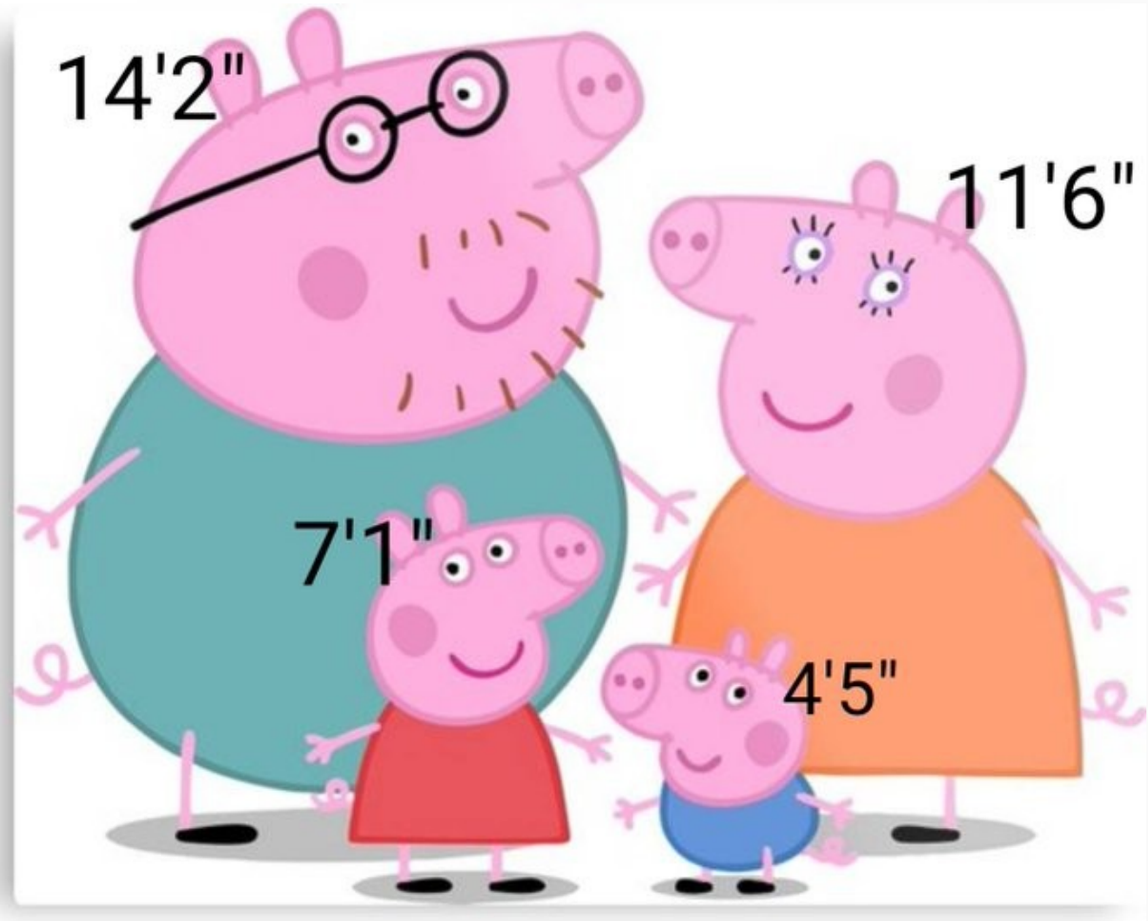
- Запретить сложные модели
- Посмотреть внутрь моделей
- Спросить у самой модели (важность)
- Суррогатные модели — например, LIME
- SHAP, или научный подход к делёжке
- Свои подходы для нейронных сетей
- Специальные «объяснимые» модели

Например, рост ребенка

Чей рост важнее?

- Средний
- Мама
- Папа

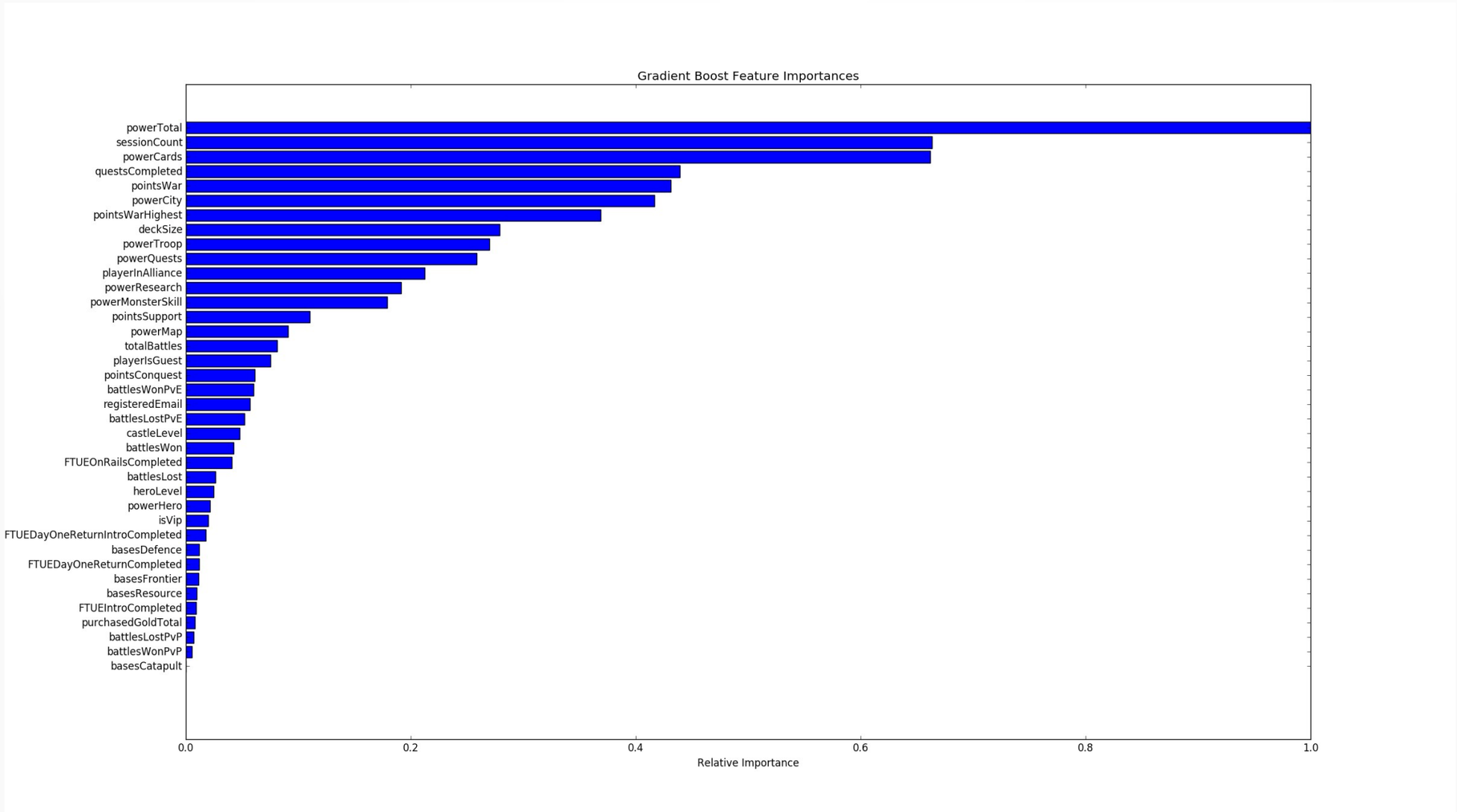
Гамильтон, регрессия



https://twitter.com/thefemalenatsu/status/1153018836335173633

Внутри модели

Важность для модели

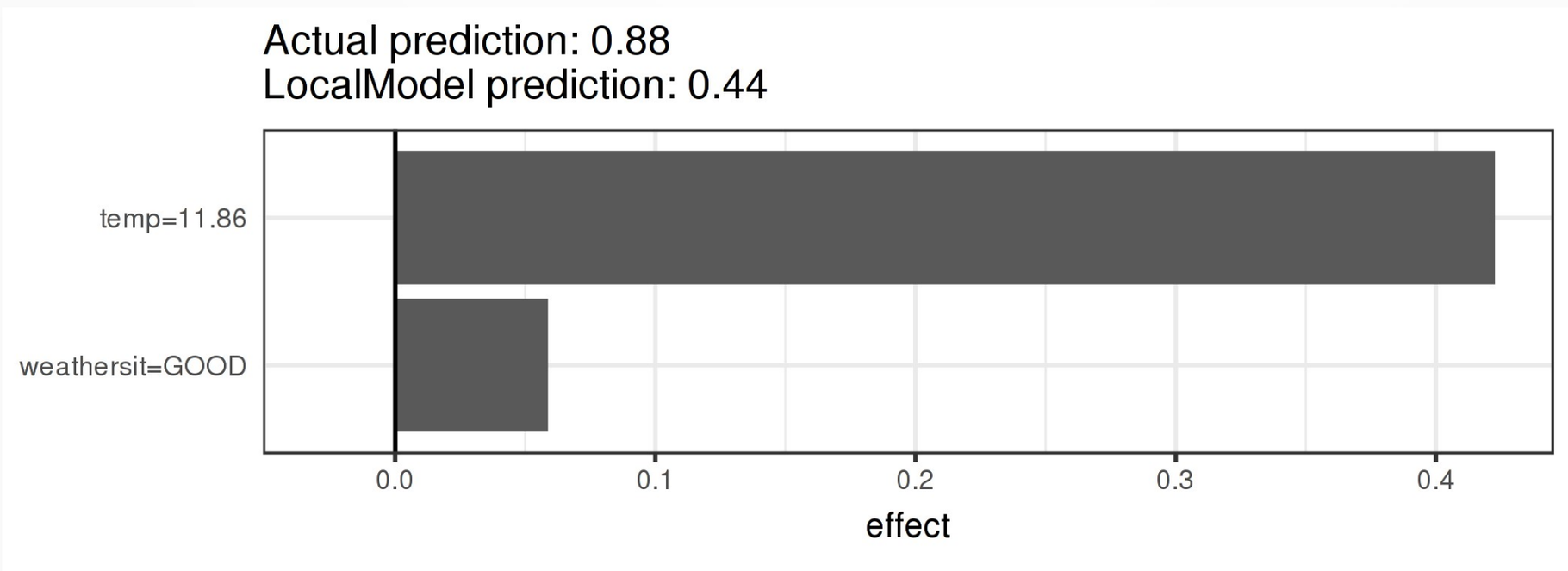


Проблемы «важности»

- Все считают по-разному
- Насколько признак был удобен для модели
- Неустойчива (при переобучении меняется)
- Неинформативна (что изменится, если...)
- Нелокальна (впрочем, «объектная важность»)
- Само по себе неинтерпретируемо
- CatBoost молодец

LIME — локальные суррогаты

- Локальная сурогатная модель
- Интерпретируемая селективная модель
- Информативна для небольших изменений

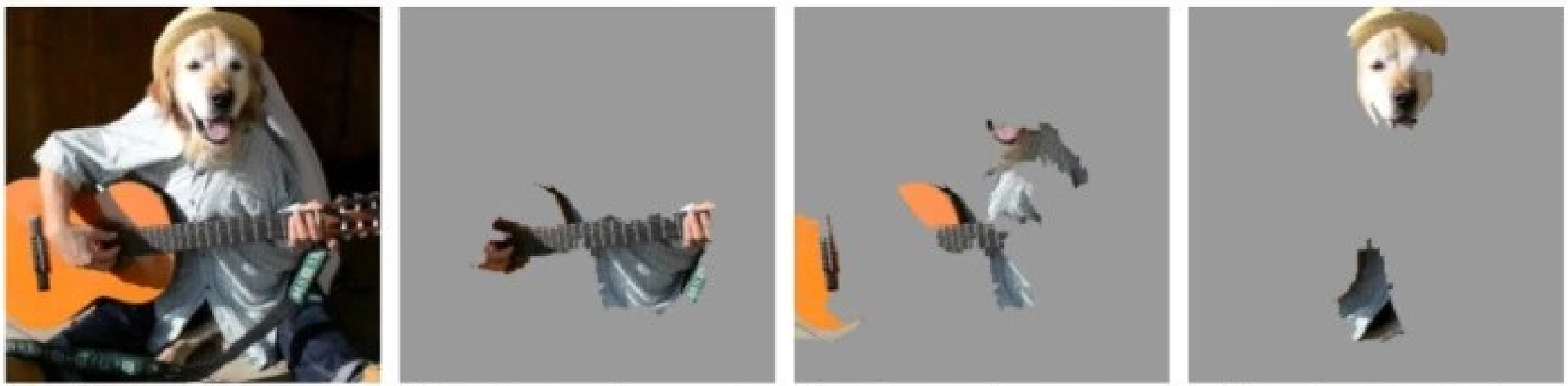


LIME — текст

- 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

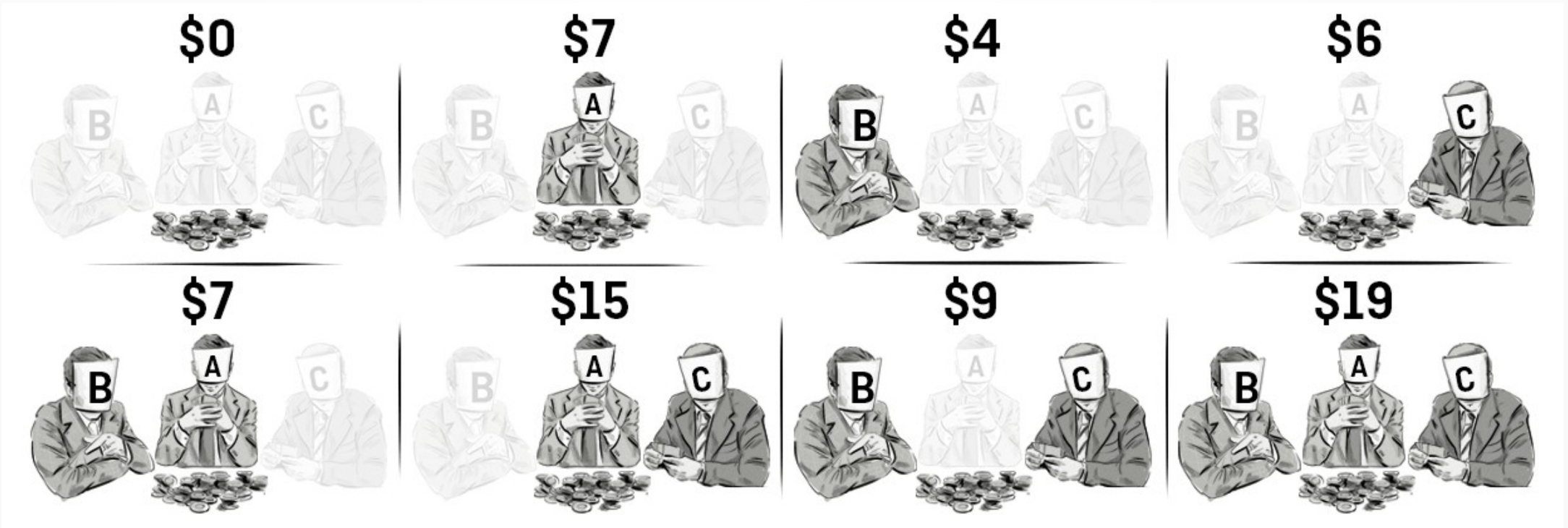
LIME — суперпиксели



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

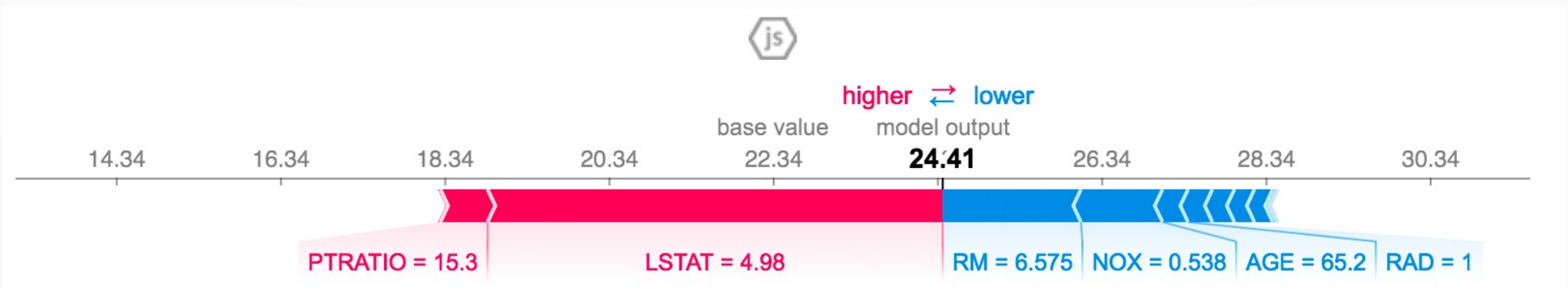
It's not a Lab in the picture, it's a Golden Retriever (a light colored one)

Shapley Values

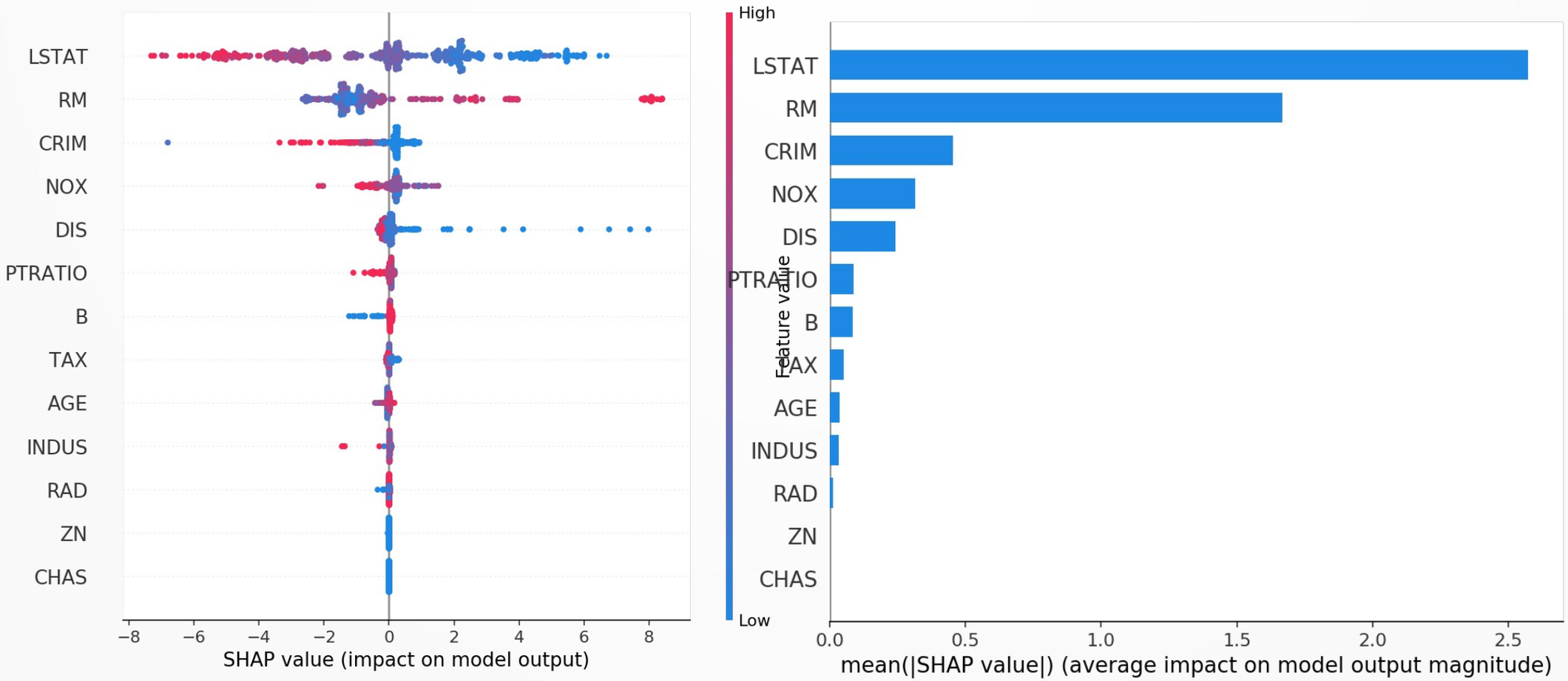


https://clearcode.cc/blog/game-theory-attribution/

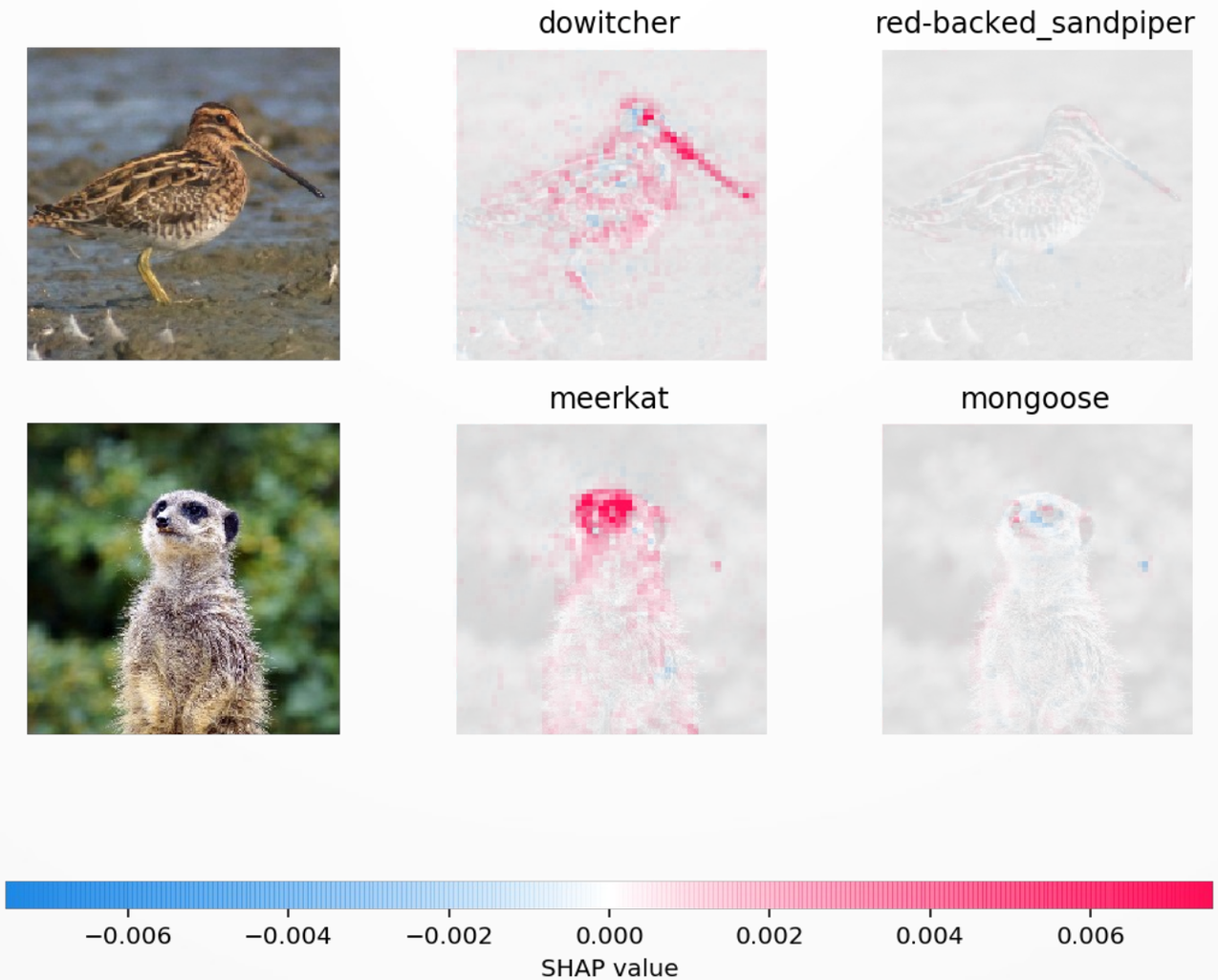
Shapley Values & SHAP



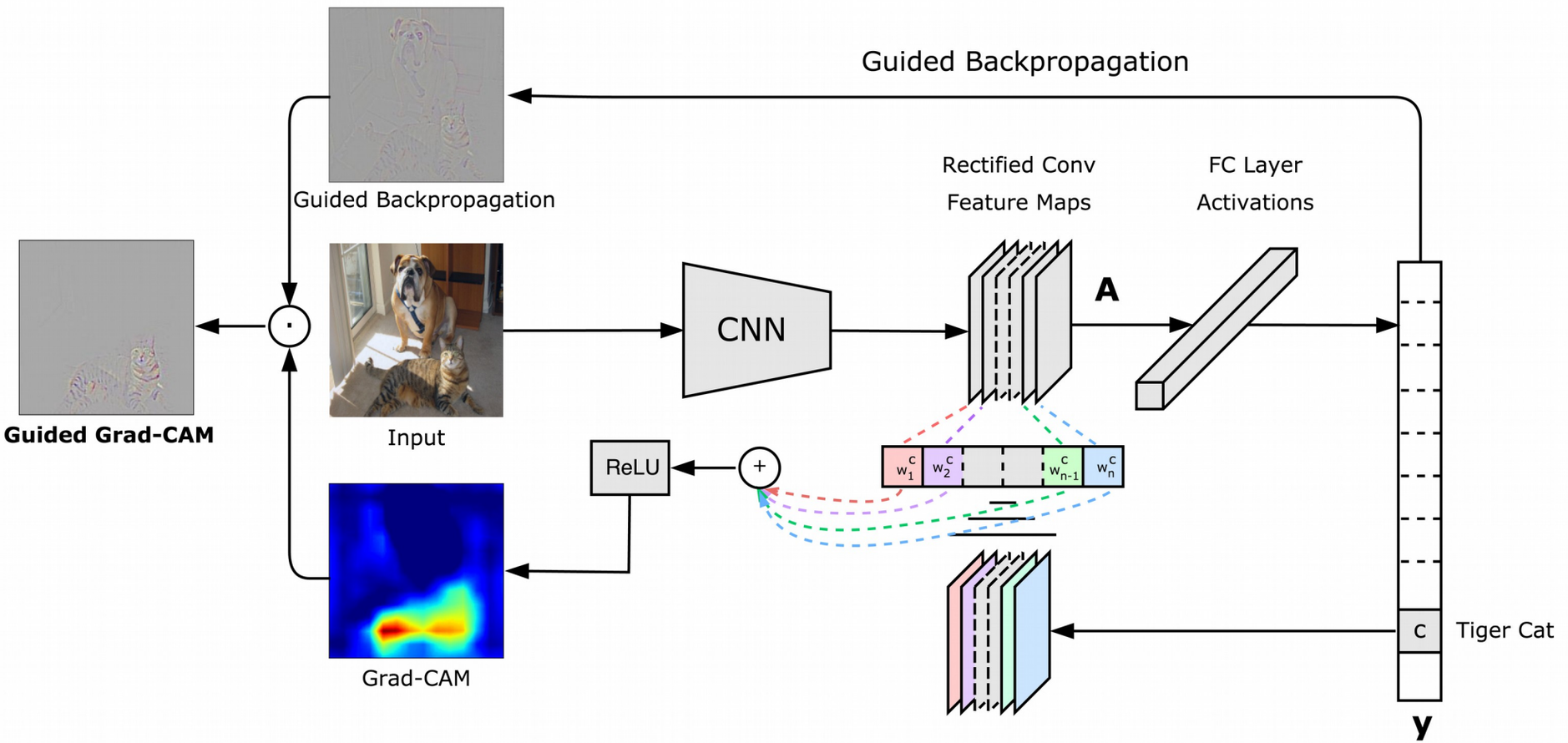
SHAP — глобальный



SHAP — влияющие точки

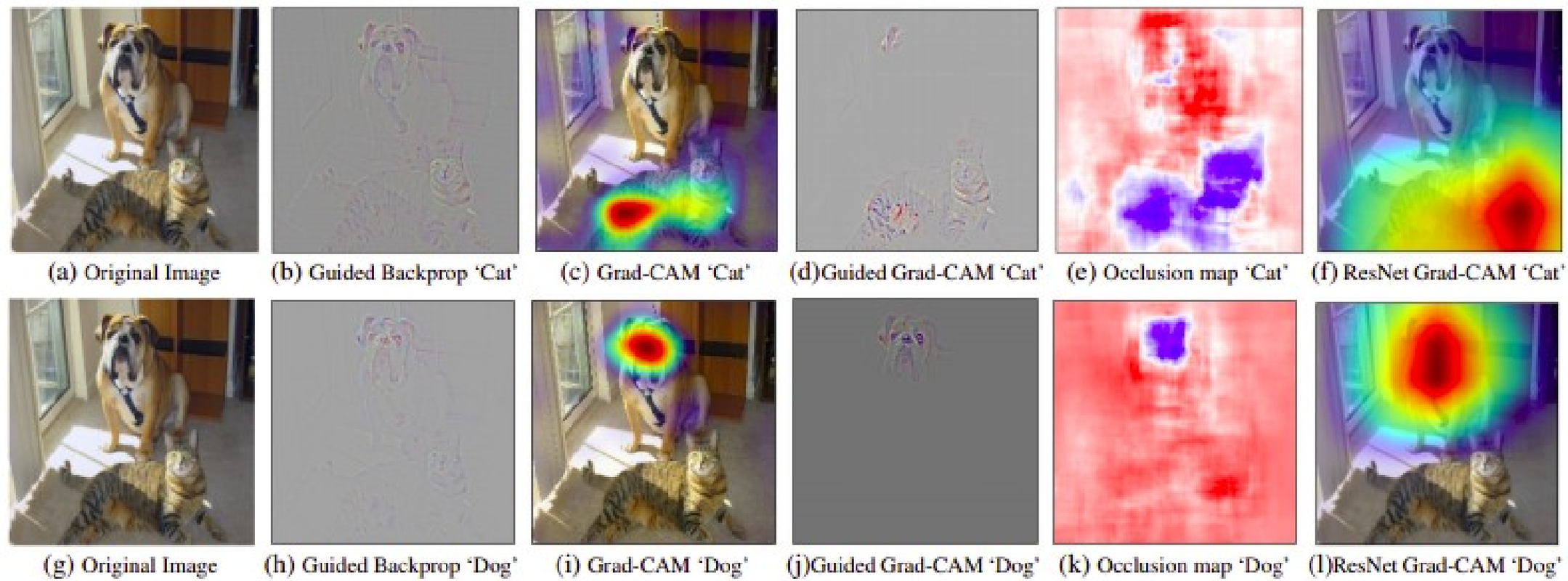


Grad-CAM



http://gradcam.cloudcv.org/

Grad-CAM



https://arxiv.org/abs/1610.02391

История картинок

- 2013 — карты значимости Saliency maps
- 2018 — карты не работают
  - Saliency map должна зависеть от весов сетки, но нет
  - Saliency map должна зависеть от закономерностей, которые есть в данных, но нет
- 2017 — карты активации нейронов
- 2020 — Concept Activation Vectors
- 2017 — Network Dissection, если примеры есть
- 2016 — Синтез примеров для объяснений

Отсюда несколько моралей

- Ученые не читают работ друг друга
- Все украдено до нас
- Для картинок обычно хватает Grad-CAM
- Для табличек обычно хватает SHAP
- Your mileage may vary

Инструменты

- SHAP
- LIME
- TreeInterpreter
- Alibi
- Anchor
- interpret

Почитать и посмотреть

- Дьяконов, Интерпретации чёрных-ящиков
- Becker, Machine Learning Explainability
- Molnar, Interpretable Machine Learning
- Google Explainable AI
- Model interpretability in Azure
- Toward Trustworthy AI Development, arXiv
- Stop Explaining BB Models, arXiv

Вопросы

Слайды тут

dkolodezev

promsoft

dkolodezev

d_key

dmitry_kolodezev

https://kolodezev.ru/download/slides-sborka-2020.pdf