Интерпретация моделей машинного обучения

Дмитрий Колодезев
ООО Промсофт, Новосибирск

CompTech@Nsk-2020 03.02.2020

План

- Как работает воркшоп
- Jupyter notebook, Google Colab и все-все-все
- Зачем все это
- Какие модели интерпретируемы
- Как можно объяснить работу модели
- Болит на сердце рана
- LIME
- SHAP
- Grad-CAM и наконец-то котики
- Ссылки и вопросы

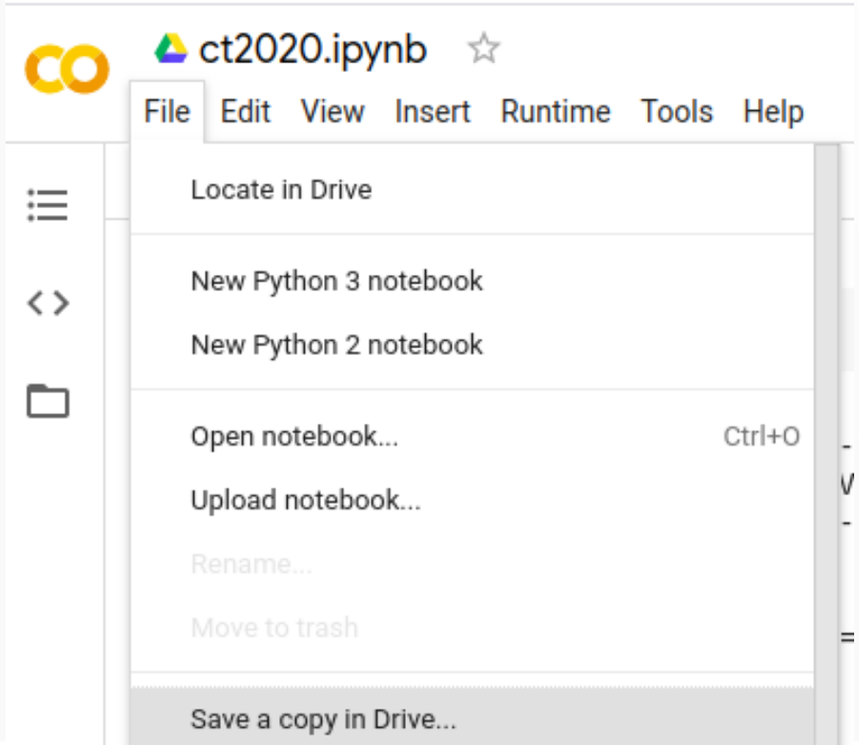
Jupyter, Colab, Халява

- Что почитать
- colab.research.google.com
- https://bit.ly/2OoKii6 Краткое введение
- https://github.com/deepmipt/dlschl/wiki
- https://habr.com/ru/post/428117/

- Здесь я открываю Colab и что-то рисую
  https://bit.ly/37P3JZ6

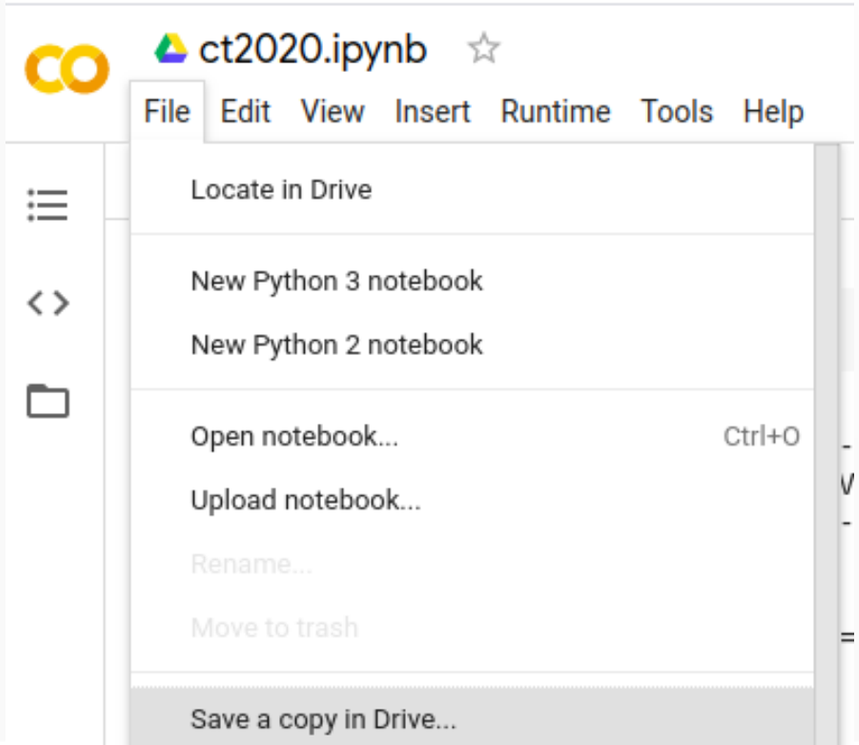
Давайте запустим котиков

- https://bit.ly/2Ol6NUO
- Сохраняем себе
- Запускаем всё

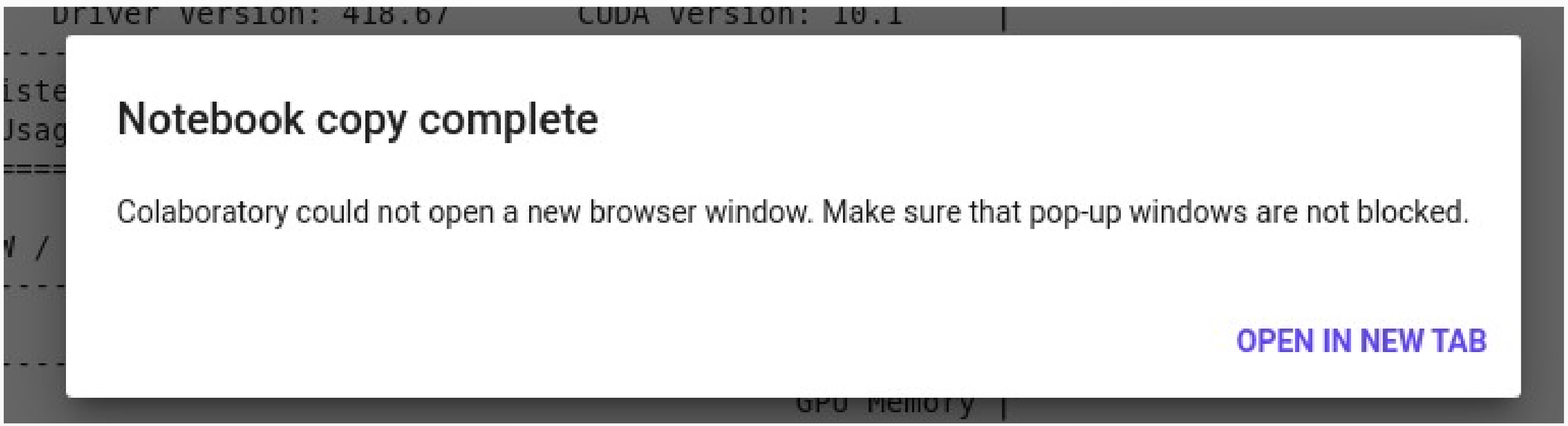


Давайте запустим кардио

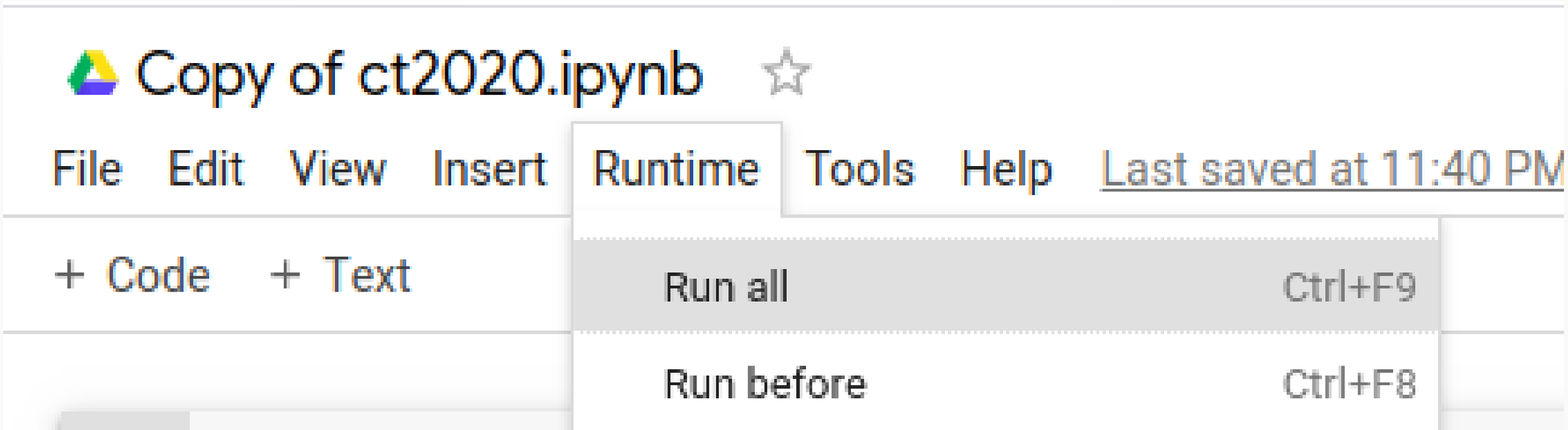
- https://bit.ly/2v0ox1i
- Сохраняем себе
- Запускаем всё



Откроем в другой вкладке



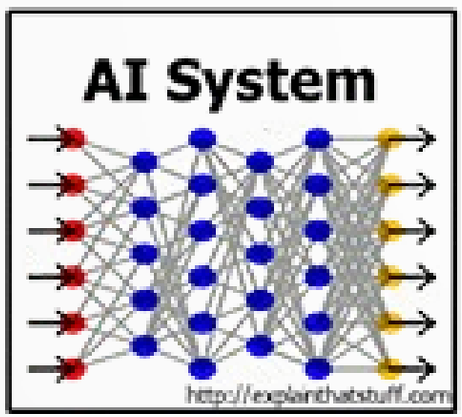
Запустим и поговорим



Осторожно, роботы!

Осторожно, люди

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

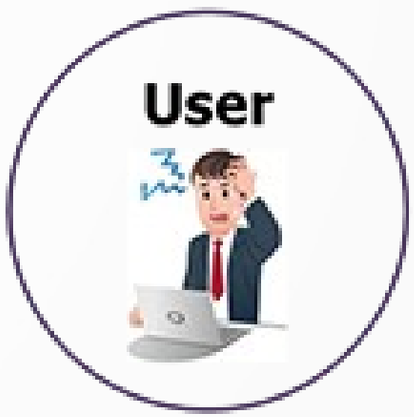
Security

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)

Что посеешь, то и пожнешь

- Обучим модель находить на видео людей с оружием (грабителей)
- Люди с оружием с камер видеонаблюдения
- Люди без оружия — из Инстаграмма
- Чему научится модель:
  - ~~Отличать грабителей~~
  - Отличать фото с камеры видеонаблюдения и фото с iPhone

Модель ищет где проще

- Пневмоторакс
- Дренаж
- На что смотрит нейронка?

Кому это нужно

- Инженерам,
  которые разрабатывают модели
- Бизнесу,
  который ими пользуется
- Конечным пользователям,
  которым с этим жить
- Регуляторам,
  которым не всё равно

Когда не нужно

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

Можно попробовать

- Изучаем данные:
  на чем училась, что мешало
- Изучаем результат:
  где ошибается, с чем справляется
- Объясняем на примерах:
  показываем характерные точки
- Документируем внутренности:
  распечатываем веса, деревья
- Создаем суррогатную модель:
  делаем модель модели

Типичные случаи (S.L.)

- Линейные модели
  - Примерно до 5 признаков интерпретируемы
- Деревья и ансамбли деревьев
  - Примерно до 3-х уровней понятны
- Списки правил
  - Так же понятны, как бухгалтерия
- kNN и их друзья
  - Молодцы

Надеюсь, оно досчиталось

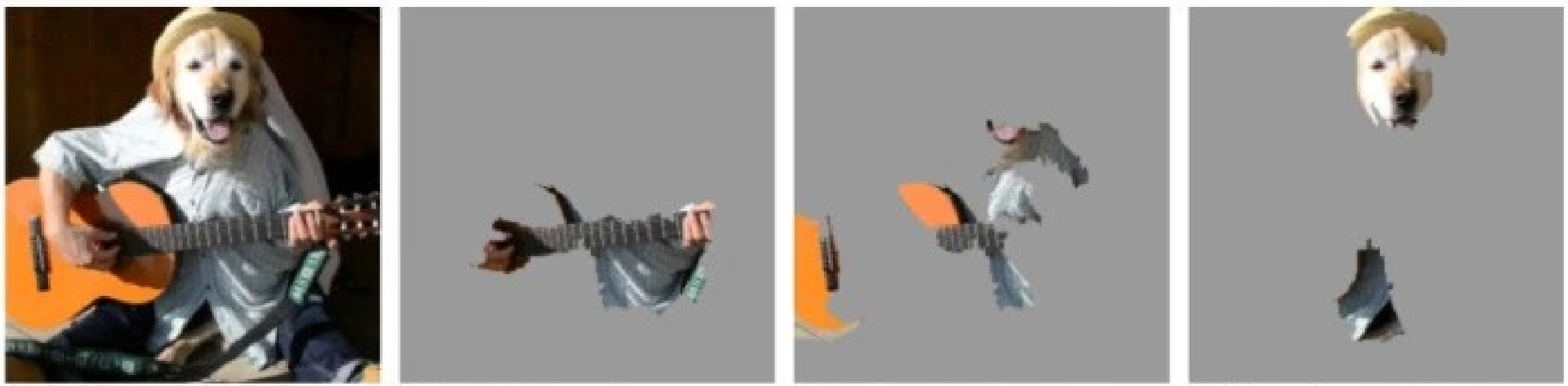
- Здесь я открываю Colab
- Обучаю случайный лес
- Объясняю все про LIME
- Все объясняю

LIME — текст

- For Christmas Song visit my channel! ;)

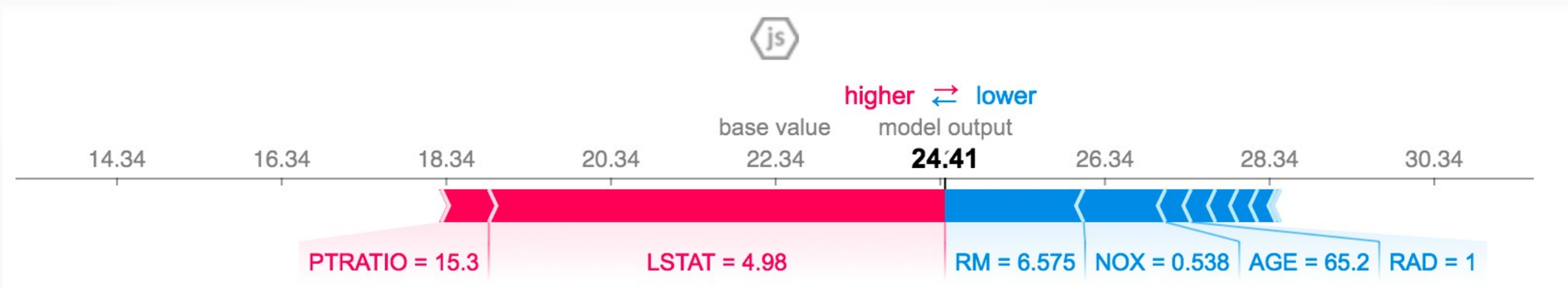
case	label_prob	feature	feature_weight
1	0.1701170	good	0.000000
1	0.1701170	PSY	0.000000
1	0.1701170	a	0.000000
2	0.9939024	channel!	6.180747
2	0.9939024	Song	0.000000
2	0.9939024	Christmas	0.000000

LIME — суперпиксели

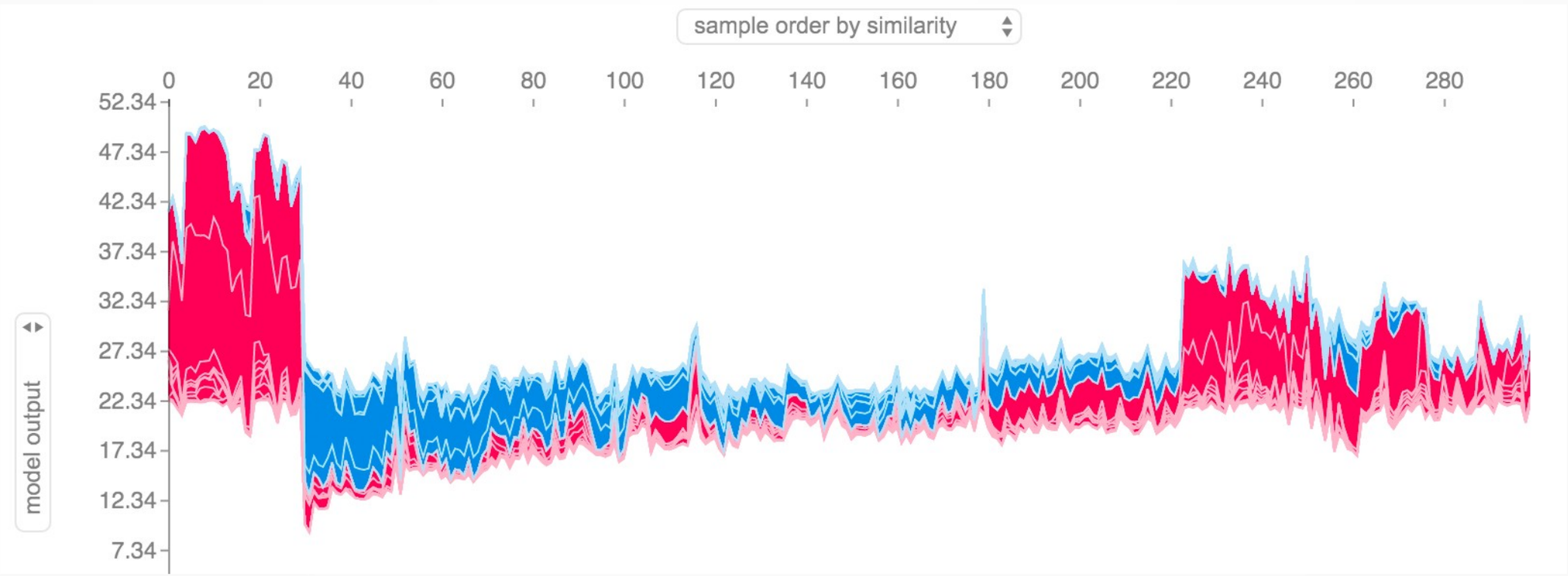


(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

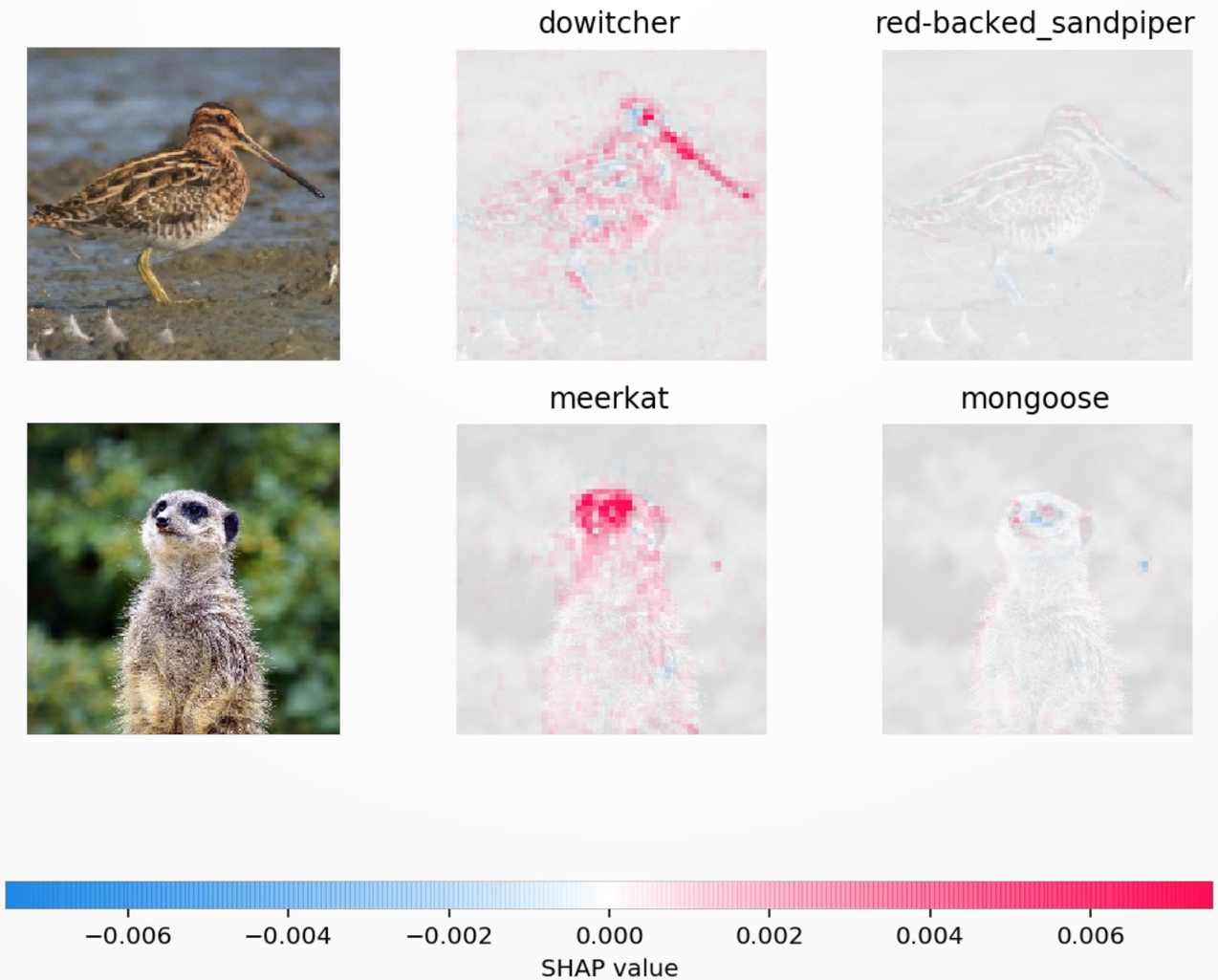
Shapley Values & SHAP



SHAP и структура данных



SHAP — влияющие точки



А теперь - котики

- Градиент пиксела ~ необычность картинки
- Grad-CAM и рассказ на DF6
- Здесь я открываю Colab и учу нейронку
- Используем FAST.AI — горячо рекомендую
- Grad-CAM в ней есть из коробки

Почитать

- Дьяконов, Интерпретации чёрных-ящиков
- Becker, Machine Learning Explainability
- Molnar, Interpretable Machine Learning
- CVPR 2018 Tutorial
- ICCV 2019 Tutorial
- MIT Network Dissection

Вопросы?

Слайды тут

dkolodezev

promsoft

dkolodezev

d_key

dmitry_kolodezev

https://kolodezev.ru/download/slides-interpretation-v4.pdf