



Прикладные задачи анализа данных

**ИНТЕРПРЕТАЦИИ ДАННЫХ И
МОДЕЛЕЙ**

**DATA INTERPRETATION
BLACK BOX INTERPRETATION**

Дьяконов А.Г.

**Московский государственный университет
имени М.В. Ломоносова (Москва, Россия)**

Интерпретация

– точного определения нет...

Zachary C. Lipton «The Mythos of Model Interpretability» 2017 <https://arxiv.org/abs/1606.03490>

Doshi-Velez & Kim Towards «A Rigorous Science of Interpretable Machine Learning» 2017

<https://arxiv.org/abs/1702.08608>

**– истолкование (данных / модели / визуализации)
с целью понимания (их / её) смысла / причин / зависимостей**

Дальше

- **интерпретация данных**
- **интерпретация моделей**

Интерпретация данных

- **важная составляющая работы DS**
 - элемент EDA
 - входит в «story telling»
- **строго говоря «обречена на ошибки...»**

Интерпретация данных

Пример неверной интерпретации...

	гв	модель	цвет	поломка
0	2015	AR20	белый	0
1	2016	NV	белый	1
2	2016	AR20	белый	0
3	2014	NV	красный	1
4	2015	Z-80	оранжевый	0
5	2015	Z-80	оранжевый	0

...

исследуем отдельные признаки

```
X.groupby('цвет')['поломка'].mean()
```

	цвет	mean_поломка
0	оранжевый	0.08
1	красный	0.12
2	чёрный	0.18
3	белый	0.20

...

Интерпретация данных

1



ПОКУПАЕТЕ ПОДЕРЖАННУЮ МАШИНУ? БЕРИТЕ ОРАНЖЕВУЮ!

Исследовательский стартап Kaggle **изучил** базу данных по покупкам подержанных автомобилей, куда также была включена информация о последующих технических проблемах с приобретенными машинами. Оказалось, что меньше проблем у тех, кто покупает автомобиль необычного цвета. Почему? Гипотеза Kaggle: цвет сигнализирует о том, что машина средство не только передвижения, но и самовыражения. О таком предмете, имеющем как утилитарное, так и символическое значение, первоначальный владелец заботится лучше.

Во всяком случае, так уверяет американская компания Kaggle: те покупатели, которые приобрели ярко-оранжевую подержанную машину, с техническими проблемами будут сталкиваться в два раза меньше, чем те, которые на цвет внимания не обратили. Интернет-платформа Kaggle не занимается продажей автомобилей. Она специализируется на объекте, который может перевернуть мир с ног на голову, — на **больших данных**. Впервые термин «**Big Data**» был использован для описания неимоверного роста цифровой вселенной.



20 неожиданных открытий, сделанных благодаря анализу данных
<http://slon.ru/specials/data-economics/articles/20-unexpected-discoveries/>

Большие данные: новый облик человечества
<http://ichip.ru/bolshie-dannye-novyjj-oblik-chelovechestva.html>

Интерпретация данных

	гв	модель	цвет	производитель	поломка
0	2015	AR20	белый	хороший	0
1	2016	NV	белый	плохой	1
2	2016	AR20	белый	хороший	0
3	2014	NV	оранжевый	плохой	1
4	2015	Z-80	оранжевый	хороший	0
5	2015	Z-80	оранжевый	хороший	0

	производитель	цвет	mean_поломка
0	хороший	оранжевый	0.08
1	хороший	красный	0.12
2	хороший	чёрный	0.09
3	хороший	белый	0.12
4	плохой	чёрный	0.22
5	плохой	белый	0.25

**Рассмотрим
производителей:
«хороший»,
«плохой»,
...**

**они выпускают машины
разных цветов в разной
пропорции...**

Интерпретация данных

**Не всегда предоставленные данные говорят
о причинах тех или иных значений**

**«надо искать только зависимости,
а их причины часто найти и обосновать невозможно»**

**В. Майер-Шенбергер, К. Кукьер Большие данные: Революция, которая изменит то, как мы живем,
работаем и мыслим // Изд-во Манн, Иванов и Фербер, 2013 г.**

Виноват ли производитель?

Интерпретация данных

**Не всегда предоставленные данные говорят
о причинах тех или иных значений**

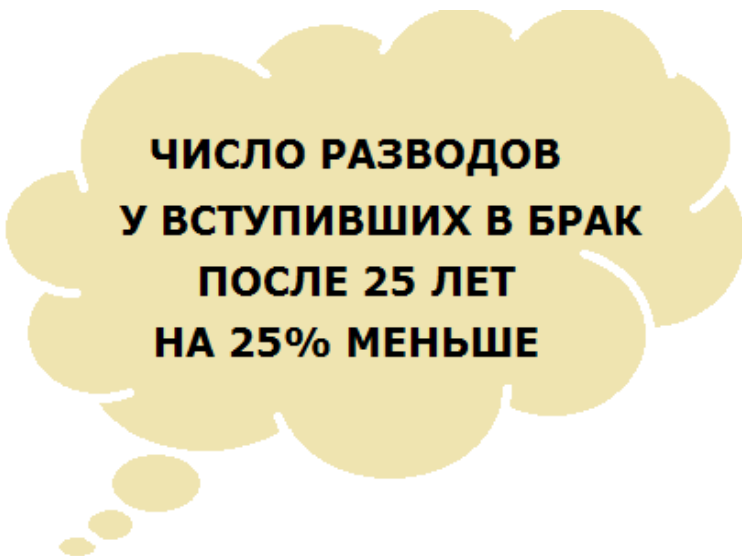
**«надо искать только зависимости,
а их причины часто найти и обосновать невозможно»**

В. Майер-Шенбергер, К. Кукьер Большие данные: Революция, которая изменит то, как мы живем, работаем и мыслим // Изд-во Манн, Иванов и Фербер, 2013 г.

Виноват ли производитель?

- м.б. не все поломки «хорошего» представлены
- м.б. значения не статистически значимы
- поломка «пороговая величина» – зависит от порога

Статистические факты ничего не значат!



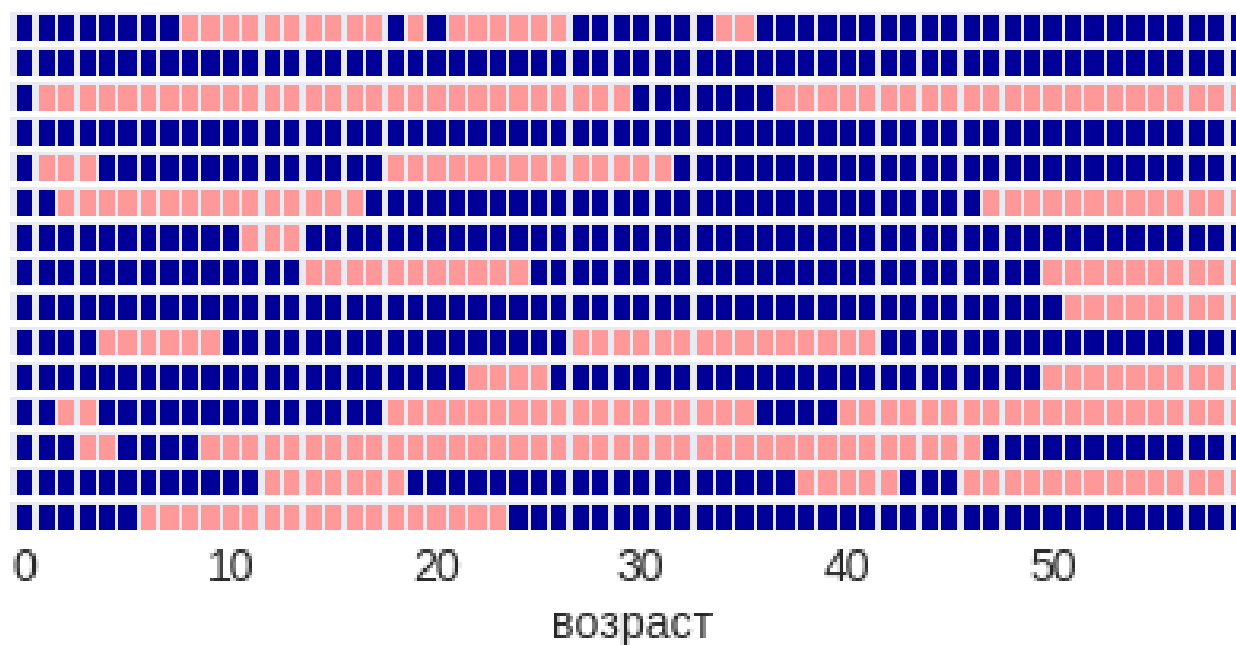
**ЧИСЛО РАЗВОДОВ
У ВСТУПИВШИХ В БРАК
ПОСЛЕ 25 ЛЕТ
НА 25% МЕНЬШЕ**

Возможные неправильные(?) интерпретации

- **взрослые люди более ответственные**
- **вторые браки более крепкие**

**Объективно есть только данные,
а выводы могут быть и необъективны!**

Обойдёмся без данных!



продолжительность жизни = 60 лет
вероятность смены статуса за год = 0.05

84.8% разводятся – из тех, кто женился
63.6% разводятся – из тех, кто женился после 25

Разница объясняется устройством мира!

Интерпретация модели

Зачем?



История о добавках...

Проблемы

- **Одна метрика качества не описывает поведения модели**

- **Использование ML в «критических областях»**

медицина, криминалистика и юриспруденция, финансы, политика, транспорт

- **Безопасность, доверия, регуляция**

ех: регулирование в финансовой сфере

ех: доверие беспилотным автомобилям

- **Второй вопрос для (И)И после «ЧТО?» – «ПОЧЕМУ?»**

Что предсказано?

источник знаний – данные

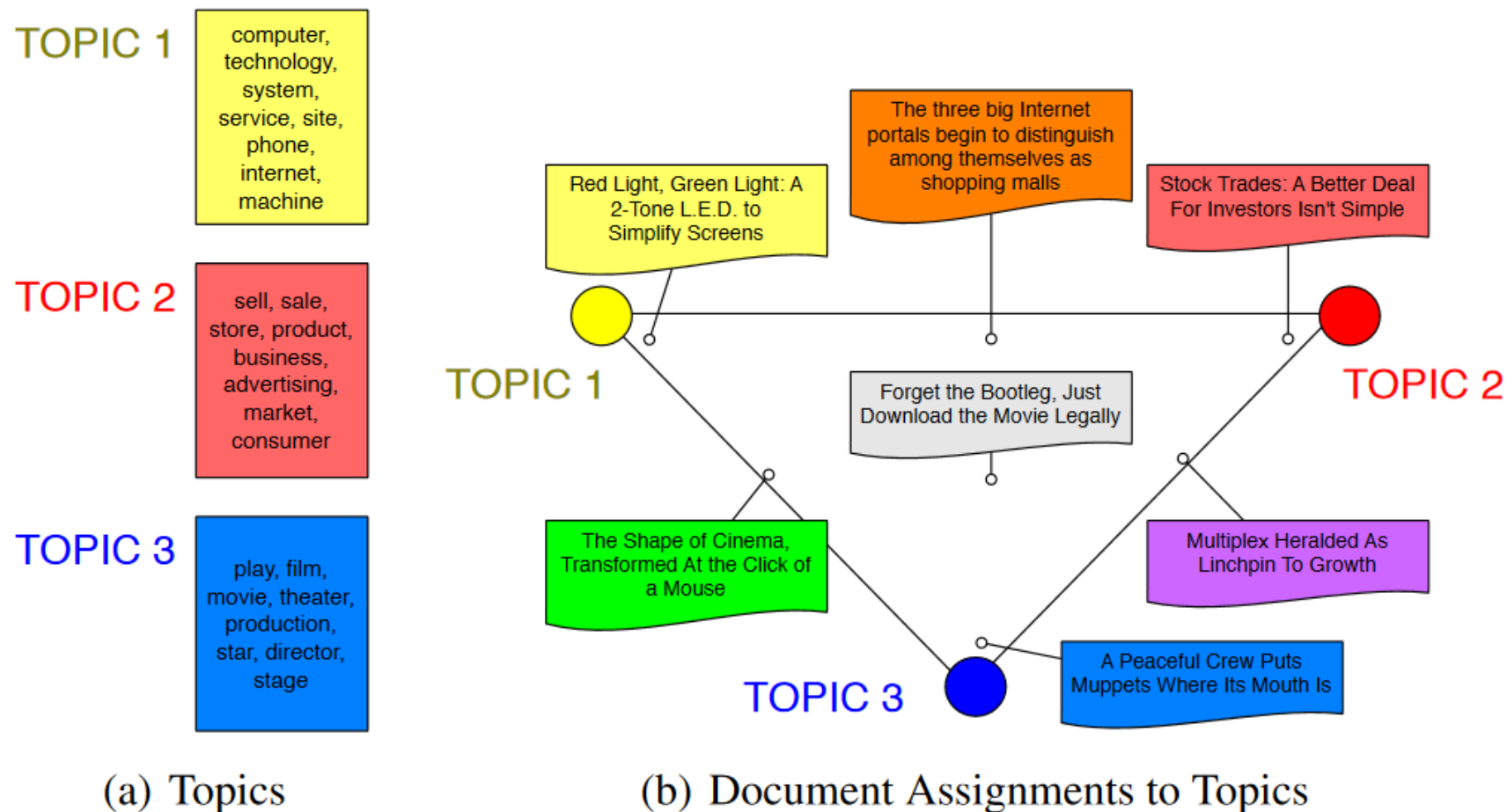
Почему предсказано?

источник знаний – модель

Способы

- **визуализация**
- **текст**
- **числа, таблицы**
- **объект(ы) / признаки / части данных**
по ним: статистики, описания, визуализации
- **аналитическая формула / простая модель**

Пример визуализации работы модели



Chang, Jonathan, Gerrish, Sean, Wang, Chong, Boyd-Graber, Jordan L, and Blei, David M. Reading tea leaves: How humans interpret topic models. In NIPS, 2009

<https://pdfs.semanticscholar.org/3a99/da22b1658695d95a764169e030cc40e2fb95.pdf>

Примеры текстовых интерпретаций

Beer (Beeradvocate)					Musical instruments (Amazon)				
pale ales	lambics	dark beers	spices	wheat beer	drums	strings	wind	microphones	software
ipa	funk	chocolate	pumpkin	wheat	cartridge	guitar	reeds	mic	software
pine	brett	coffee	nutmeg	yellow	sticks	violin	harmonica	microphone	interface
grapefruit	saison	black	corn	straw	strings	strap	cream	stand	midi
citrus	vinegar	dark	cinnamon	pilsner	snare	neck	reed	mics	windows
ipas	raspberry	roasted	pie	summer	stylus	capo	harp	wireless	drivers
piney	lambic	stout	cheap	pale	cymbals	tune	fog	microphones	inputs
citrusy	barnyard	bourbon	bud	lager	mute	guitars	mouthpiece	condenser	usb
floral	funky	tan	water	banana	heads	picks	bruce	battery	computer
hoppy	tart	porter	macro	coriander	these	bridge	harmonicas	filter	mp3
dipa	raspberries	vanilla	adjunct	pils	daddario	tuner	harps	stands	program

Интерпретация тем из

McAuley, Julian and Leskovec, Jure. Hidden factors and hidden topics: understanding rating dimensions with review text. In RecSys. ACM, 2013. <https://cs.stanford.edu/people/jure/pubs/reviews-recsys13.pdf>

Пример – зачем

**При классификации текстов Christianity / Atheism
качество 94%, но конкретные классификации основаны
на частоте слов «re», «posting», «host»
т.е. дело в специфике набора данных!**

<https://lilianweng.github.io/lil-log/2017/08/01/how-to-explain-the-prediction-of-a-machine-learning-model.html>

Примеры текстовых интерпретаций

*This is a **Black-Capped Vireo** because...*



Description: this bird has a white belly and breast black and white wings with a white wingbar.

Explanation-Dis: this is a bird with a white belly yellow wing and a **black head**.

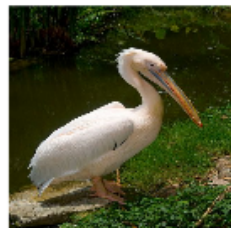
*This is a **Crested Auklet** because...*



Description: this bird is black and white in color with a orange beak and black eye rings.

Explanation-Dis.: this is a black bird with a **white eye** and an orange beak.

*This is a **White Pelican** because...*



Description: this bird is white and black in color with a long curved beak and white eye rings.

Explanation: this is a large white bird with a **long neck** and a **large orange beak**.

*This is a **Geococcyx** because...*



Description: this bird has a long black bill a white throat and a brown crown.

Explanation-Dis.: this is a black and white spotted bird with a **long tail feather** and a pointed beak.

Объяснение изображений (Visual Explanations) из

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, Trevor Darrell Generating Visual Explanations 2016 <https://arxiv.org/pdf/1603.08507.pdf>

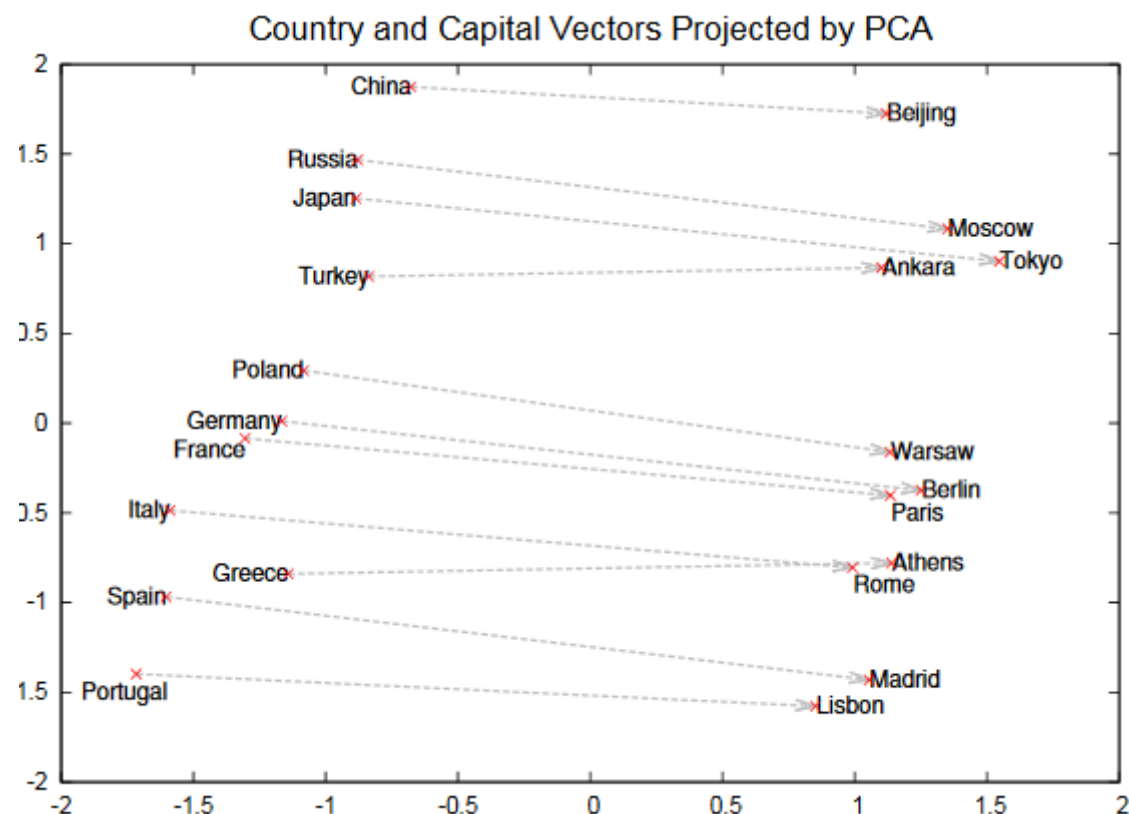
Пример интерпретаций с изображениями



<https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html>

**далее: улучшения качества работы с помощью интерпретаций
тут: особенность изображения «гантель» – она всегда в руке**

Пример интерпретации с помощью объектов



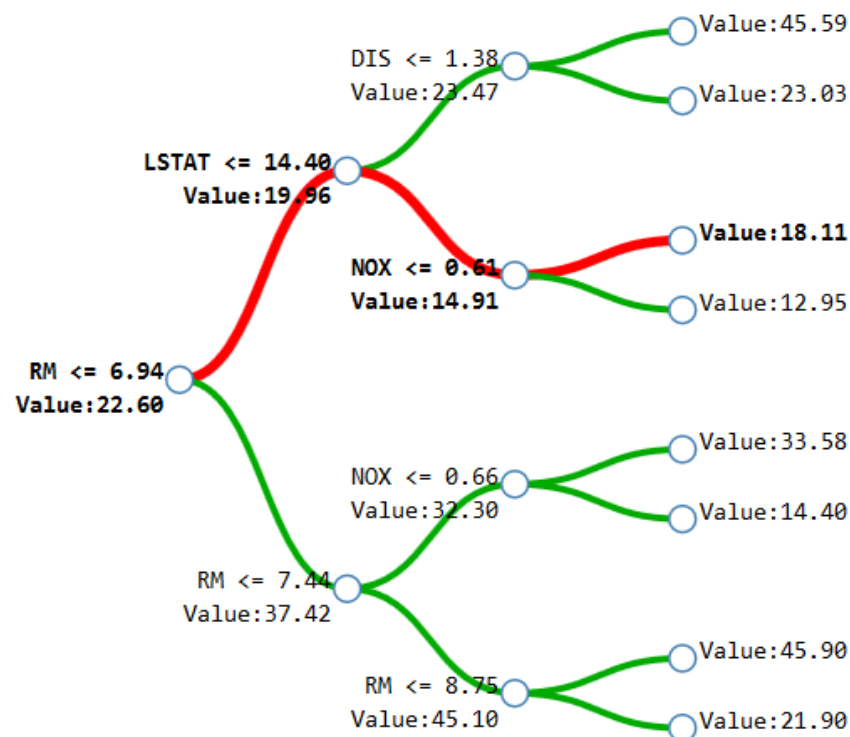
2 главные компоненты метода PCA из 1000-мерного представления слов

Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg S, and Dean, Jeff. Distributed representations of words and phrases and their compositionality. In NIPS, pp. 3111–3119, 2013. <https://arxiv.org/pdf/1606.03490.pdf>

Аналогично изображают скрытые слои нейросетей

Пример интерпретацией формулой

деревья, леса, бустинг...



Prediction: 18.11 \approx 22.60 (trainset mean) - 2.64(loss from RM) - 5.04(loss from LSTAT) + 3.20(gain from NOX)

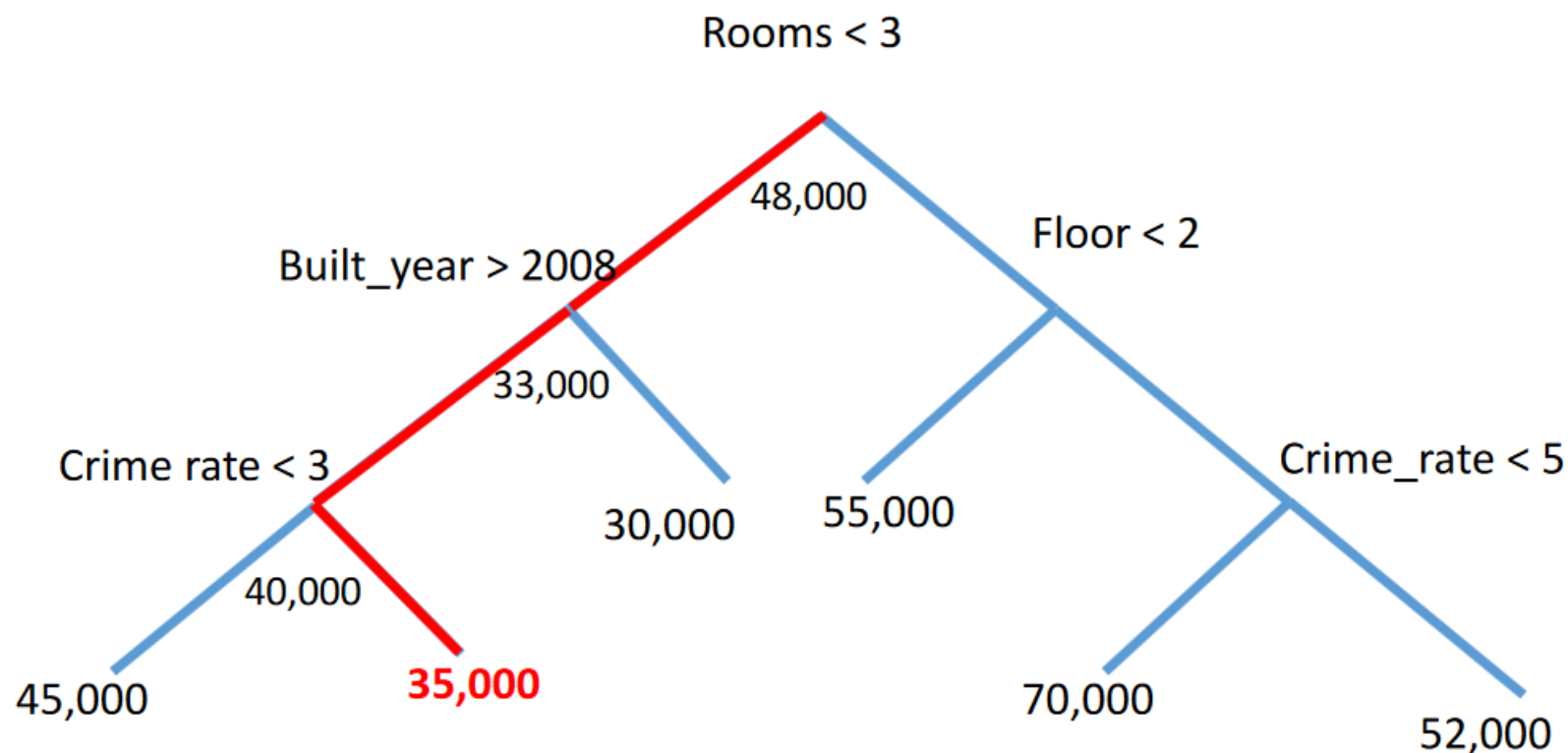
RM	LSTAT	NOX	DIST	Predict
6.5	16.1	0.12	2.2	

<http://blog.datadive.net/interpreting-random-forests/>

<https://github.com/andosa/treeinterpreter>

Пример интерпретацией формулой

**treeinterpreter – для конкретного прогноза:
среднее цели + вклад признака 1 + вклад признака 2 + ...**



$$\text{Price} = 48,000 - 15,000(\text{Rooms}) + 7,000(\text{Built_year}) - 5,000(\text{Crime_rate}) = 35,000$$

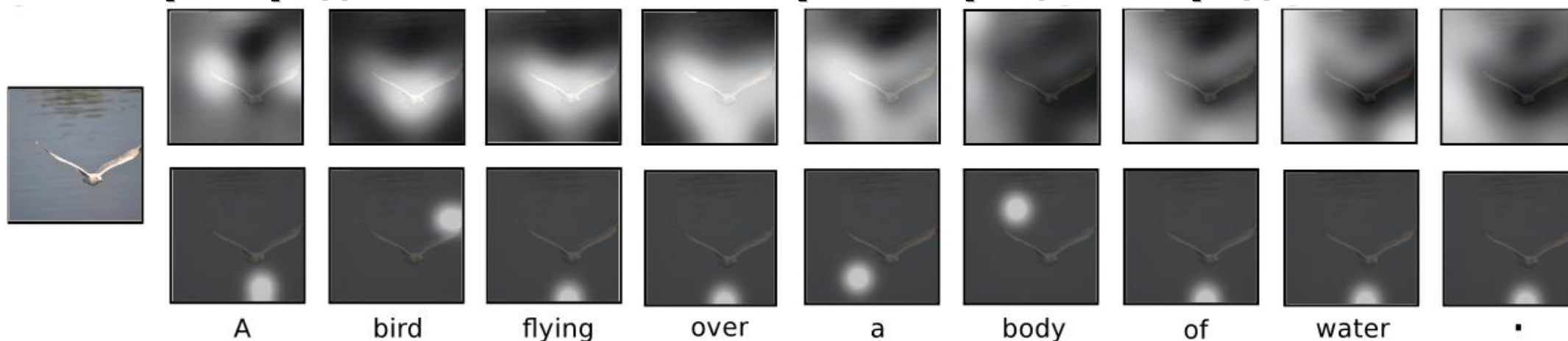
<http://cs.ioc.ee/~tarmo/tday-kao/saabas-slides.pdf>

Для чего используется интерпретация

- **объяснение результатов (самое главное)**
почему дан такой ответ (2й шаг в построении ИИ)
- **улучшения качества решения**
как увидим, сюда входит и оценка важности признаков
- **понимание, как устроены данные**
на самом деле – модель
- **проверка перед имплементацией**

Пример улучшения качества работы

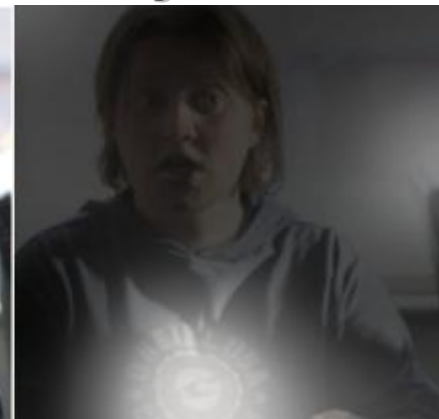
Как распределяется внимание при генерации очередного слова



Верхний ряд – большая область внимания, нижний – маленькая
Заодно инструмент «что модель видит»... – и почему ошибается...



A large white bird standing in a forest.



A woman holding a clock in her hand.

«Show, Attend and Tell: Neural Image Caption Generation with Visual Attention» [Kelvin Xu и др. 2016
<https://arxiv.org/abs/1502.03044>]

Типы интерпретации

Глобальная
объяснение работы всей модели

Локальная
объяснение конкретной ситуации (ex: отдельного прогноза)

**Часто можно сделать алгоритм прозрачным (transparency),
но не объяснить конкретное предсказание**

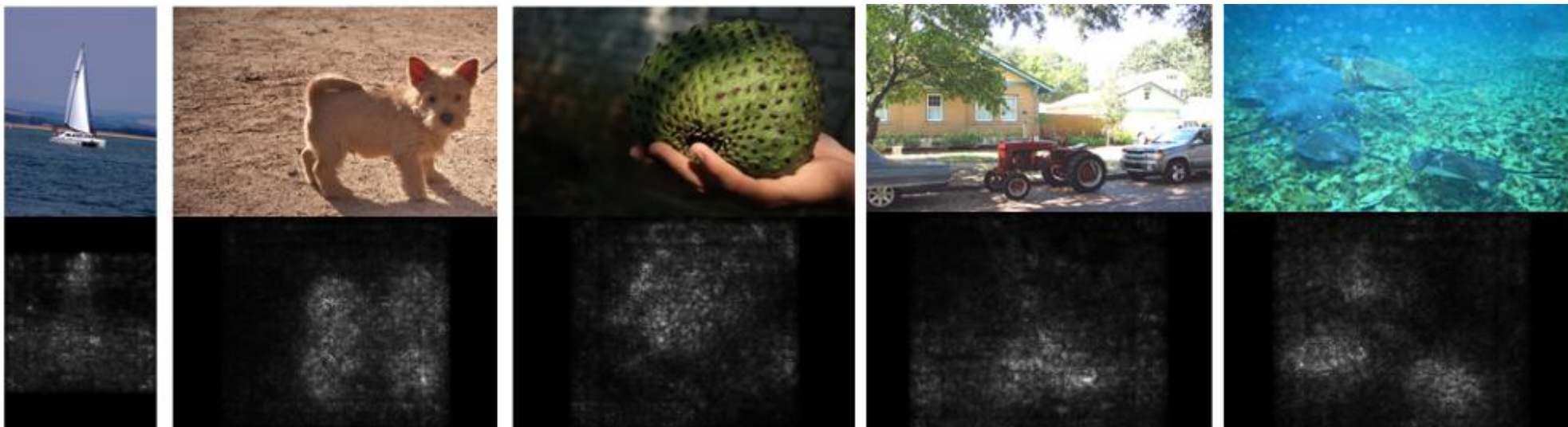
НС на каждом последующем слое производит поиск закономерностей, связанных со всё более сложными графическими примитивами (на первом слое – детектирование границ)

Отношение к модели

- **model-agnostic**

нет предположений относительно исследуемой модели

Пример локальной интерпретации



что надо изменить в объекте, чтобы максимально изменился ответ модели

Karen Simonyan, Andrea Vedaldi, Andrew Zisserman Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps 2014 <https://arxiv.org/abs/1312.6034>

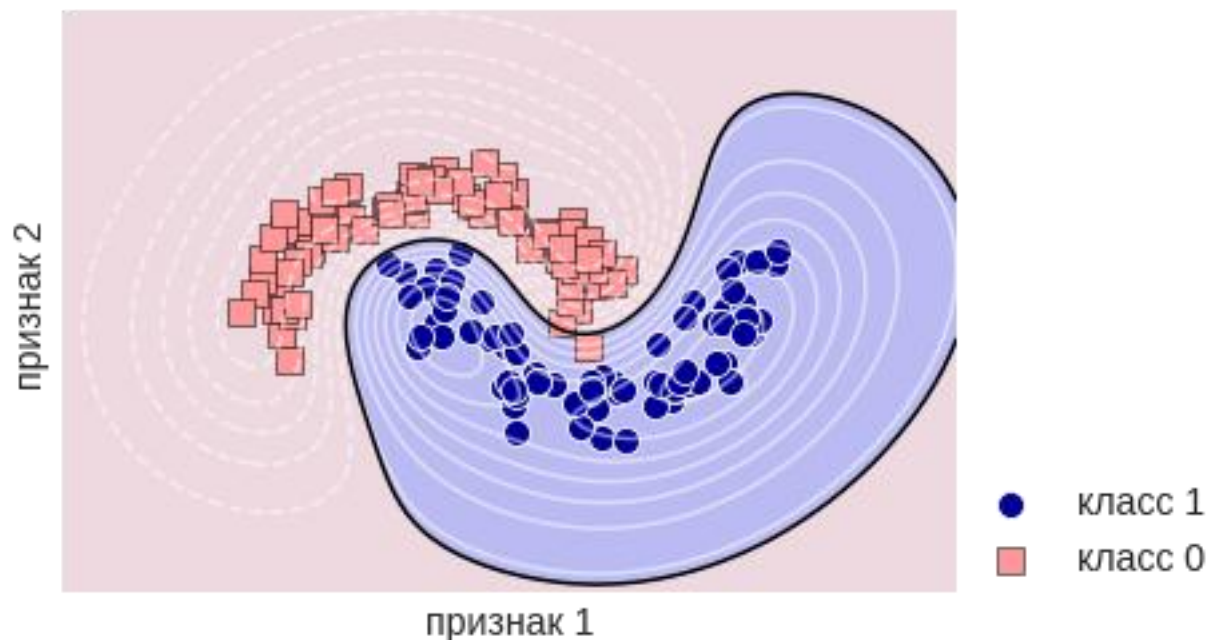
Требование к понятности интерпретации

- **сравнение / сопоставление (contrastive)**
не просто «почему не дали кредит»,
а «почему мне дали, а ему не дали»,
«что сделать, чтобы дали»
- **краткость и конкретика (выборочность / selectivity)**
чётко указать 1-3 причины
- **контекстность**
на языке клиента, учитывая предметную область
- **соответствие ожиданиям и правдивость**
объяснение работает в других ситуациях

Tim Miller «Explanation in Artificial Intelligence: Insights from the Social Sciences» 2017
<https://arxiv.org/abs/1706.07269>

Пример несоответствия ожиданиям

Разделяющая поверхность для SVM с ядром



**Может быть: в данных есть монотонность (по какому-то признаку),
а в ответах модели нет (особенность геометрии разделения)**

**ех: «Если бы Ваши доходы стали отрицательны,
Вы бы получили кредит»**

Интерпретация линейной модели

$$a(X_1, \dots, X_n) = w_0 + w_1 X_1 + \dots + w_n X_n$$

- **Интерпретация слагаемых формулы**
 - **Визуализация весов**
 - **Визуализация эффектов признаков**
 -

$$\begin{array}{l} \text{СТОИМОСТЬ} \\ \text{квартиры} \end{array} = 50000 \cdot \begin{array}{l} \text{ПЛОЩАДЬ} \\ \text{КВ.М.} \end{array} + 100000 \cdot \begin{array}{l} \text{НАЛИЧИЕ} \\ \text{балкона} \end{array} + \dots$$

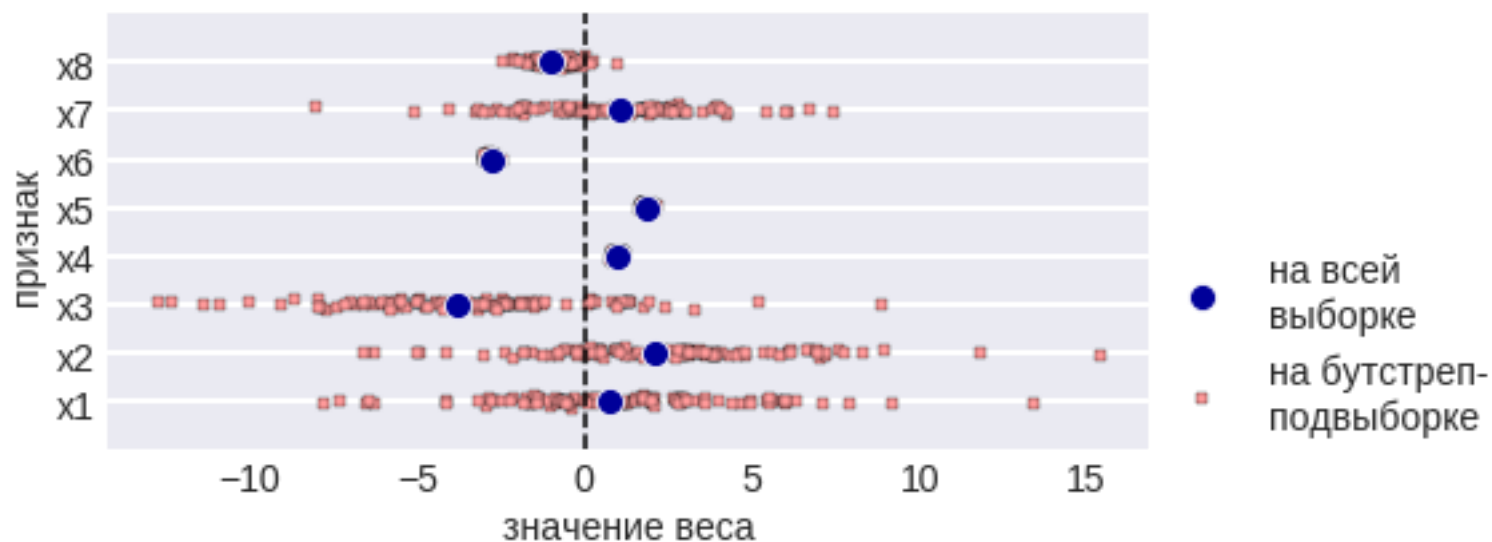
Интерпретация слагаемых формулы

«увеличение площади на 1 кв.м. увеличивает её стоимость на ...»

«наличие балкона увеличивает стоимость квартиры на ...»

в логистической регрессии влияем на отношение правдоподобия

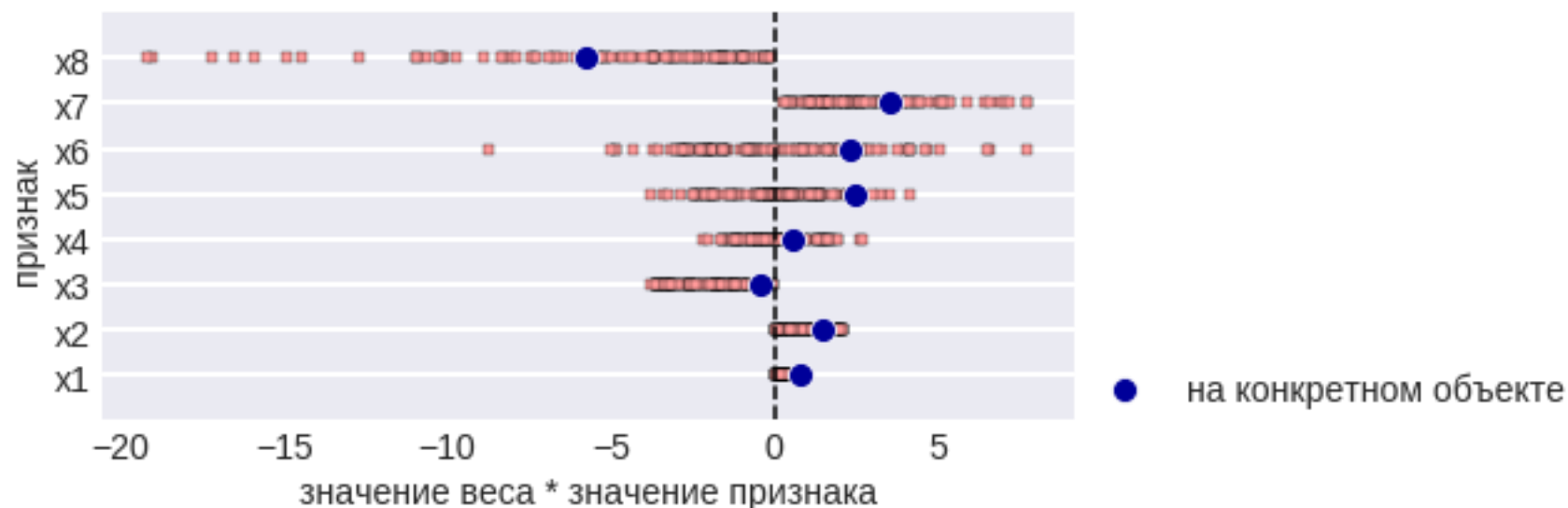
Визуализация весов (Weight Plot)



– зависит от масштаба

Визуализация эффектов признаков (Effect Plot)

эффект = вес * значение



можно использовать ящик с усами или плотности

+ нет проблемы масштаба

+ можно визуализировать отдельные объекты

Визуализация отдельных объектов

на рис. можно отметить конкретные предсказания

Интерпретация чёрных ящиков

- Анализ частичной зависимости (PDP)
 - Анализ взаимодействия признаков
- Индивидуальное условное ожидание Individual Conditional Expectation (ICE)
 - Важности признаков (Feature Importance)
- Глобальные суррогатные модели (Global Surrogate Models)
 - Локальные суррогатные модели (ex: LIME)
 - Shapley Value Explanations (SHAP)
 - Исследование отдельных блоков модели

~ плохо, когда все признаки коррелируют

Анализ частичной зависимости (Partial Dependence)

$$X_{-i} = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$$

$$PD_i(X_i) = \int a(X_1, \dots, X_n) \partial P(X_{-i})$$

$$PD_i(X) = \frac{1}{m} \sum_{(X_1, \dots, X_n) \in \text{train}} a(X_1, \dots, X_{i-1}, X, X_{i+1}, \dots, X_n)$$

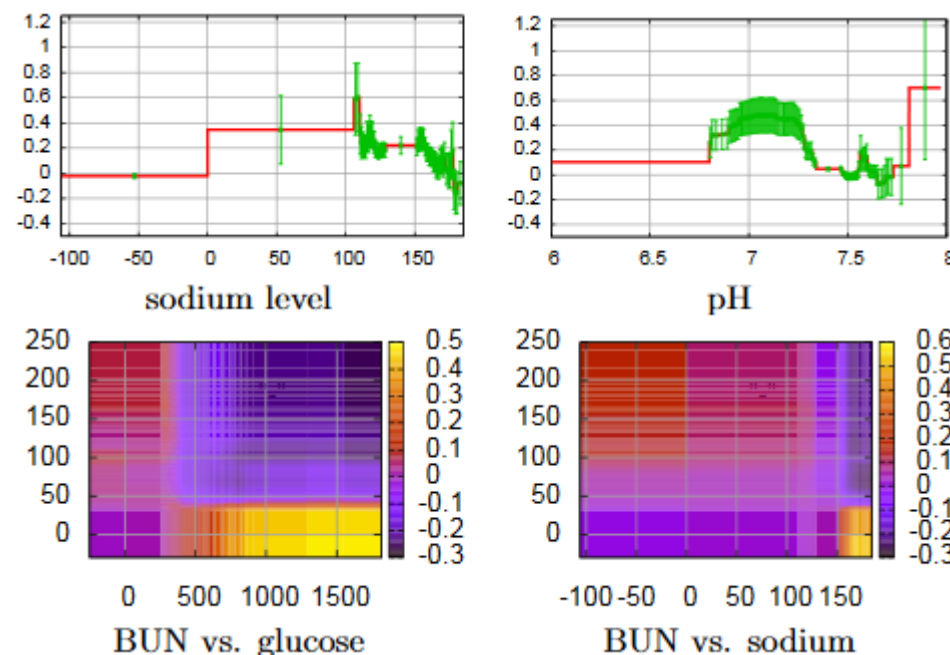
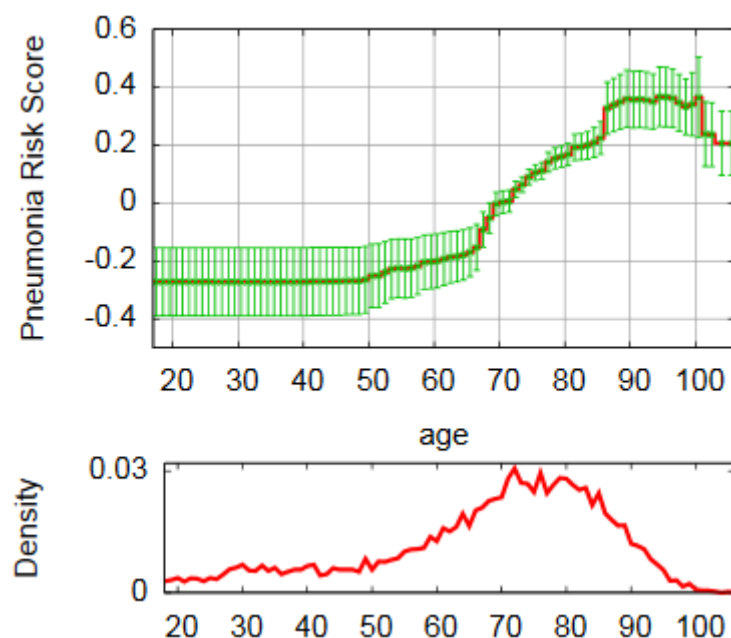
- + можно 2-3-мерные графики!**
- не учитывается распределение по признаку**
- предполагается независимость от других признаков**
 - кумулятивный эффект (см. дальше)**

Анализ частичной зависимости



PDP = Partial Dependence Plot

Пример зависимостей от признаков



зелёный коридор – при варьировании модели с помощью бэгинга

Caruana, Rich, Lou, Yin, Gehrke, Johannes, Koch, Paul, Sturm, Marc, and Elhadad, Noémie. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In KDD, 2015.

<http://people.dbmi.columbia.edu/noemie/papers/15kdd.pdf>

Анализ взаимодействия признаков

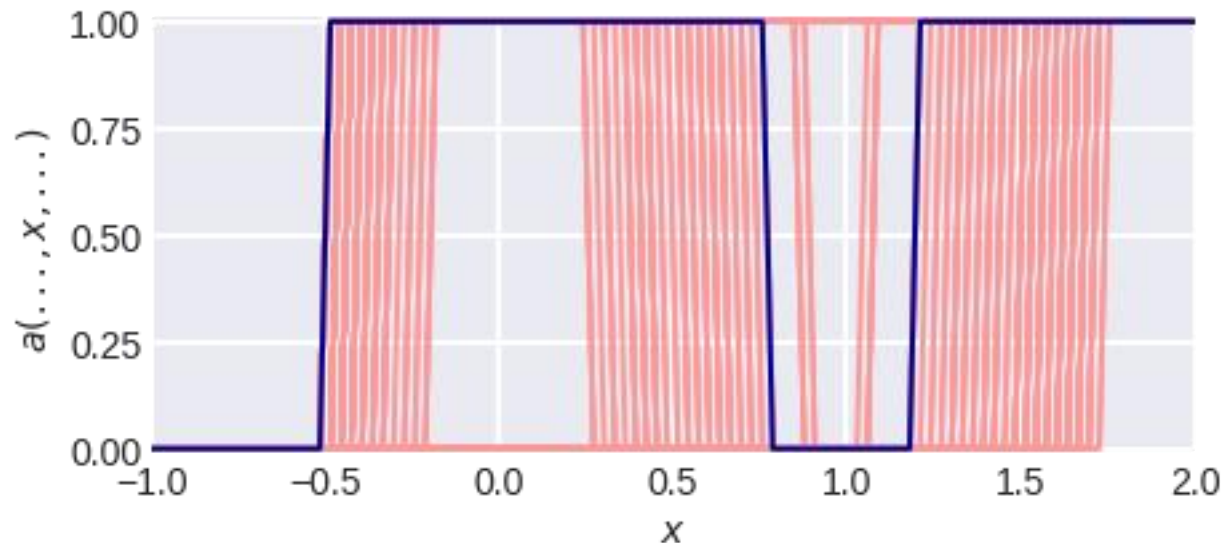
с помощью H-статистики (Friedman, Popescu)

$$H_{ij}^2 = \frac{\int PD_{ij}(X_i, X_j) - PD_i(X_i) - PD_j(X_j)}{\int PD_{ij}^2(X_i, X_j)}$$

интегрирование тут – сумма по элементам выборки

- + описывает, как много взаимодействие привносит в модель**
- + есть теория**
- + оценивает сложные взаимодействия (не определённого вида)**
- долго вычислять**
- не говорит о виде взаимодействия**

Индивидуальное условное ожидание Individual Conditional Expectation (ICE)



Одна линия – один объект
как меняется результат модели при изменении значения признака у
объекта

имеет смысл изображать не все объекты

Индивидуальное условное ожидание Individual Conditional Expectation (ICE)

PDP = усреднение ICE

- + более интуитивный (если линии хорошо видны)**
- + раскрывает некоторые зависимости**
 - строится только для 1 признака**

Важности признаков (Feature Importance)

**идея ~ как сильно уменьшится качество,
если испортить конкретный признак**

Breiman (2011); Fisher, Rudin, Dominici (2018)

портим – как правило, перемешиваем

- оценка именно для этой модели**

не надо обучать новую / и это будет некорректно

- просто реализовать**
- сохраняет распределение по значениям признака**

+ понятная интерпретация

– не всегда интересна (часто интересуется, а как признак меняет модель)

– нужны размеченные данные

Оценить ошибку модели

$$e(\text{test}) = \frac{1}{|\text{test}|} \sum_{x \in \text{test}} L(a(x), y(x))$$

**1. Для каждого t -го признака
сгенерировать обучающую выборку с перемешанным значением
этого признака:**

$$\text{test} \xrightarrow{\text{permute } t} \text{test}_t$$

важность t -го признака:

$$\frac{e(\text{test}')}{e(\text{test})} \text{ или } e(\text{test}') - e(\text{test})$$

Важности (варианты вычисления)

- перемешиванием

- **treeinterpreter** (выше)

для конкретного прогноза: смещением от среднего целевого с помощью каждого признака

- **scikit-learn**

общим уменьшением минимизируемого функционала
сумма по всем...

- **SHAP (SHapley Additive exPlanations)**

<https://github.com/slundberg/shap>

**в разных реализациях разные оценки важности
поэтому некорректно сравнивать время работы**

SHAP (SHapley Additive exPlanations)

Пусть $f(x | S)$ – значение модели, обученной на подмножестве признаков $S \subseteq \{1, 2, \dots, n\}$, на объекте x , тогда смещение, обеспечиваемое i -м признаком

$$\phi_i = \sum_{S \subseteq \{1, 2, \dots, n\} \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} (f(x | S \cup \{i\}) - f(x | S))$$

Вычисляется, например методом Монте-Карло

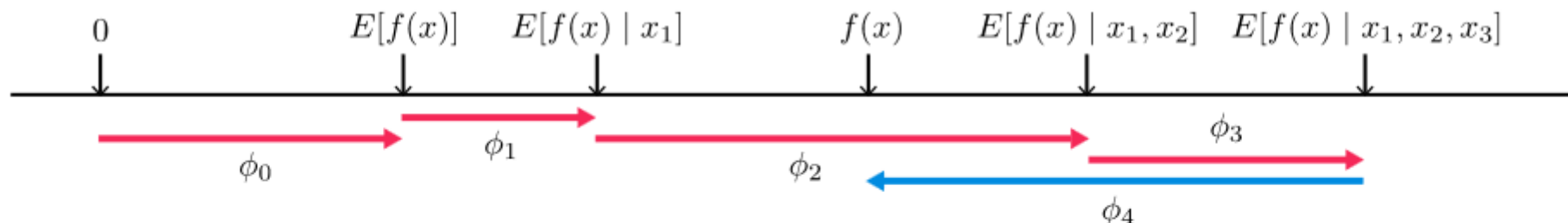


Figure 2: SHAP (SHapley Additive exPlanation) values explain the output of a function f as a sum of the effects ϕ_i of each feature being introduced into a conditional expectation. Importantly, for non-linear functions the order in which features are introduced matters. SHAP values result from averaging over all possible orderings. Proofs from game theory show this is the only possible consistent approach where $\sum_{i=0}^M \phi_i = f(x)$. In contrast, the only current individualized feature attribution method for trees satisfies the summation, but is inconsistent because it only considers a single ordering [24].

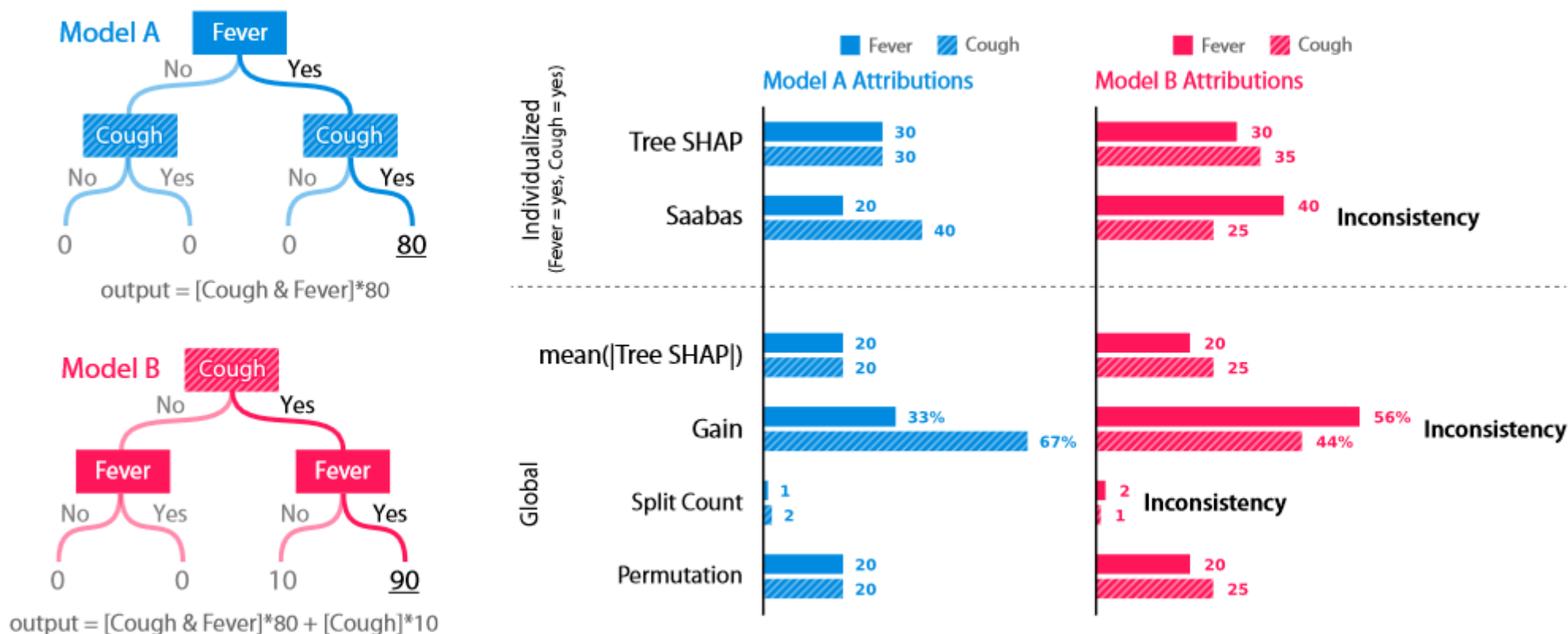
<https://arxiv.org/pdf/1706.06060.pdf>

<http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>

<https://arxiv.org/pdf/1802.03888.pdf>

Согласованность (Consistency)

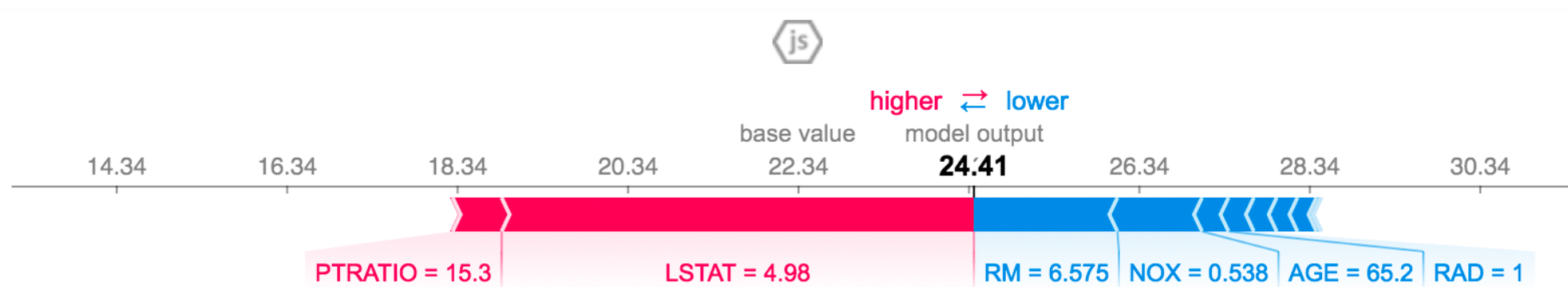
- если модель изменить так, что она более существенно начинает зависеть от какого-то признака, то его важность не убывает



все методы, кроме перемешивания и SHAP несогласованы!

<https://arxiv.org/pdf/1706.06060.pdf>

Пример SHAP-визуализации



есть базовое значение (среднее на обучении)

есть ответ модели

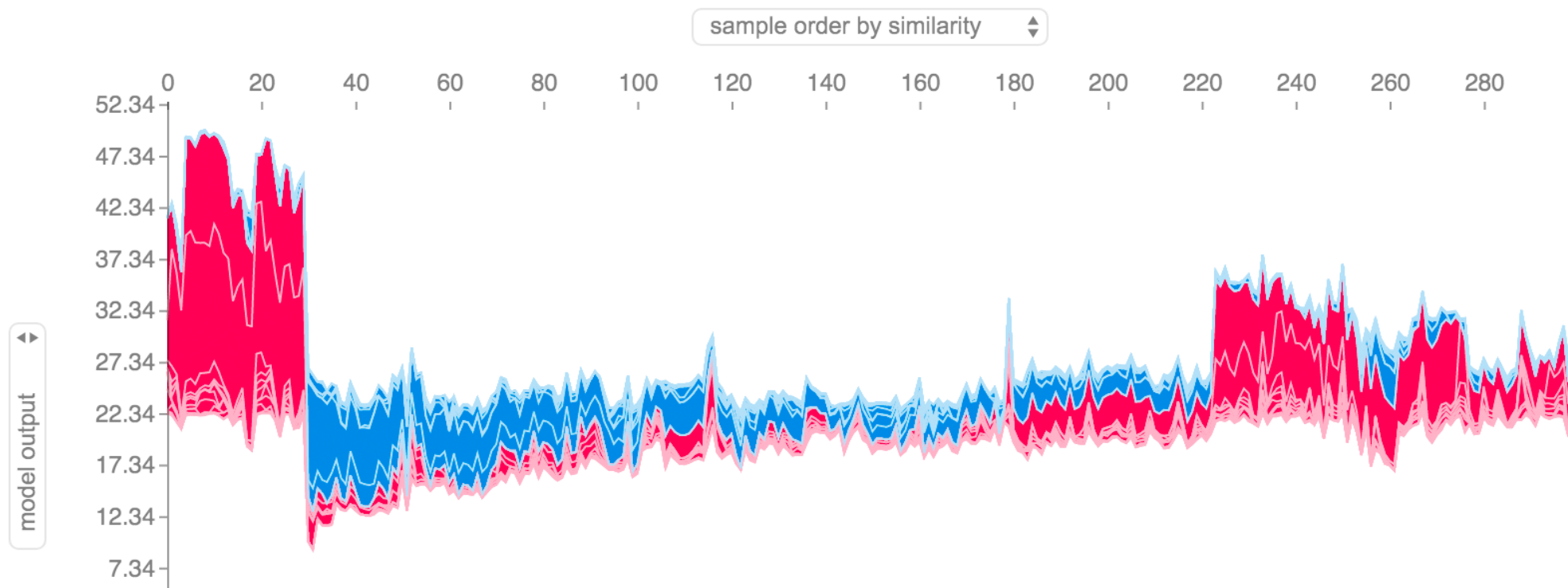
**показан вклад каждого признака в смещение от базового
(надписаны – с большим вкладом)**

красные – смещают вправо

синие – смещают влево

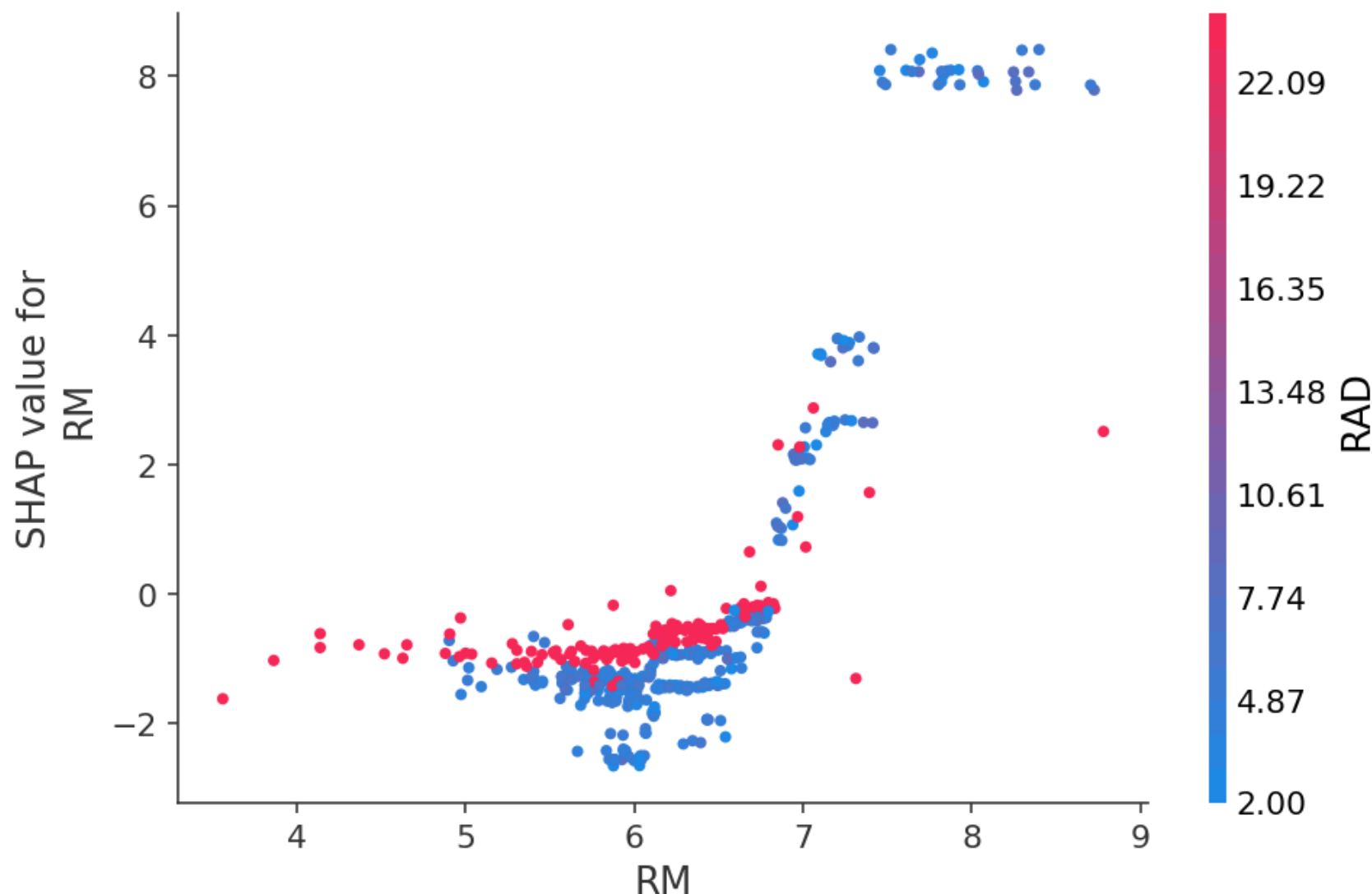
<https://github.com/slundberg/shap>

Пример SHAP-визуализации



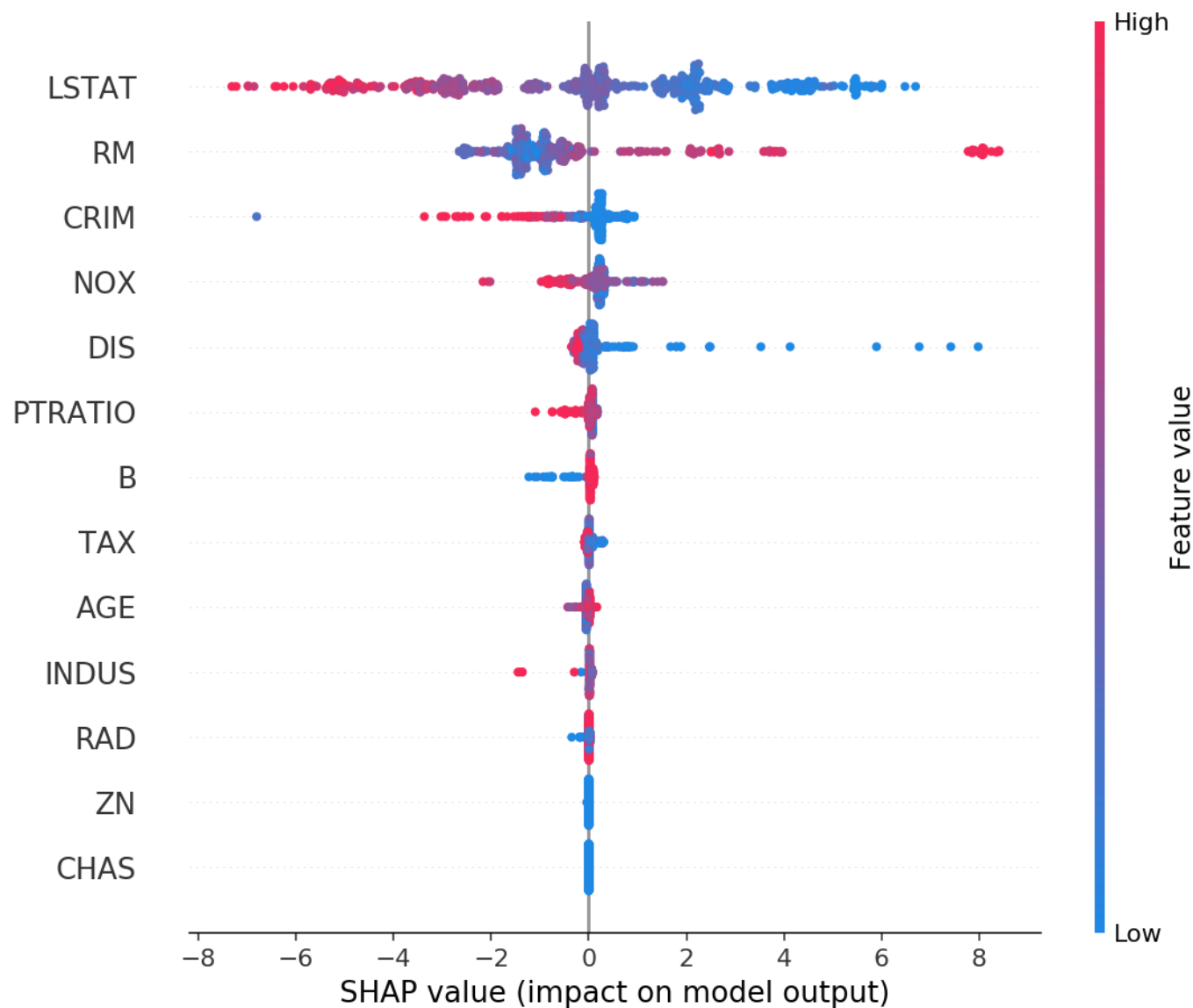
**Развернули предыдущий рисунок +
построили его на ответе каждого объекта из теста**

Пример SHAP-визуализации



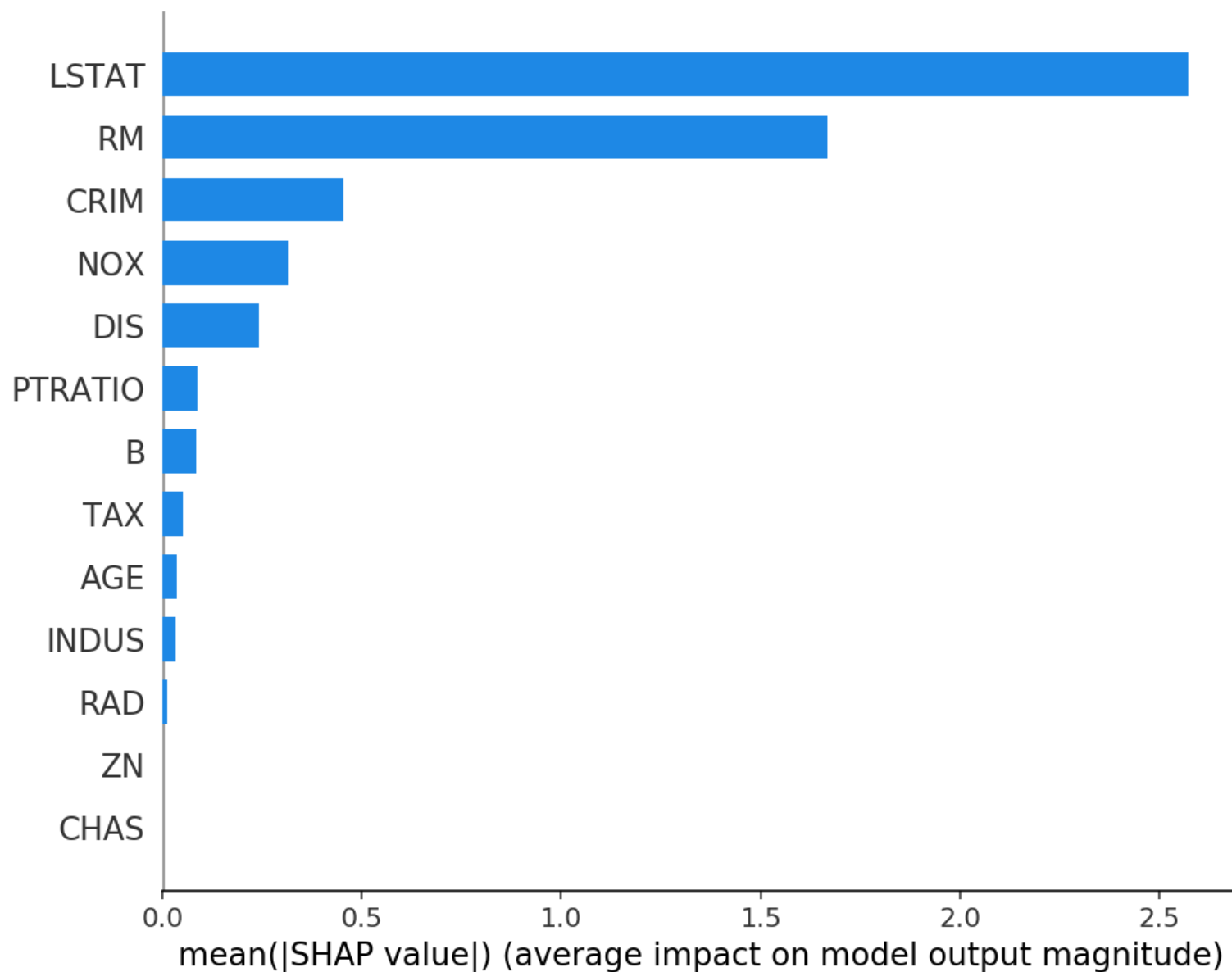
значение признака – его SHAP value (для всего датасета)
раскраска просто по другому признаку – RAD

Пример SHAP-визуализации



как отдельные признаки влияют в сравнении

Пример SHAP-визуализации



Просто средний модуль SHAP values

Пример SHAP-визуализации



Что знать про важности признаков

- **нет идеального алгоритма оценки важности признаков**
(для любого можно подобрать пример, когда он плохо работает)
- **если много похожих, то важность может «делиться между ними»**
не рекомендуется отбрасывать признаки по порогу важности
- **??? модель для решения задачи и оценка важности должны основываться на разных парадигмах**
не рекомендуется оценивать важность с помощью RF и потом настраивать его же на важных признаках

Глобальные суррогатные модели (Global Surrogate Models)

- **простая интерпретируемая модель, которая обучается, чтобы моделировать поведение чёрного ящика**

кстати, можно не иметь выборки;)

- + гибкость: можем проиграться с выбором суррогатной модели (и м.б. данными для неё)**

- **двойное обучение, накапливается ошибка (уже точно не говорим про данные)**

- **может недостаточно хорошо приблизить**

Dark Knowledge

**~ простая модель учится предсказывать ответы сложной,
ответы в широком смысле – ех: вектор вероятностей**

также термины

model compression / model distillation

(Bucila et al., 2006; Ba & Caruana, 2014; Hinton et al., 2015)

Хороший обзор

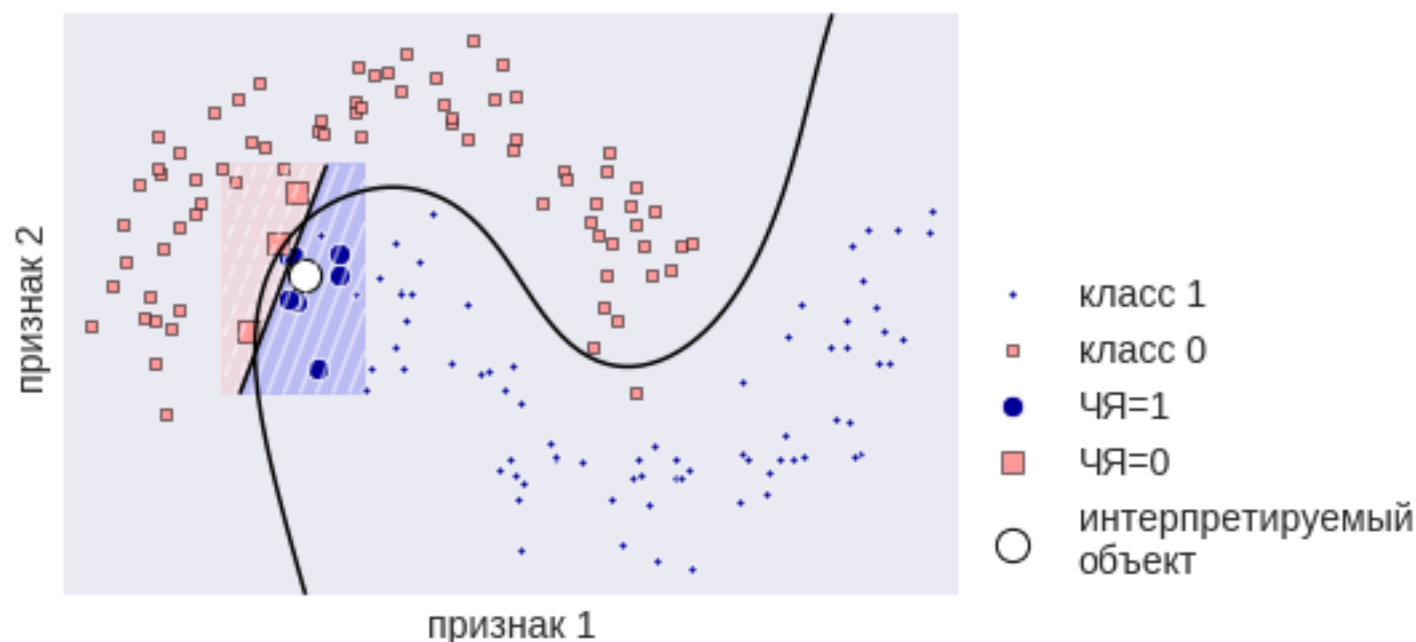
Interpreting Deep Classifiers by Visual Distillation of Dark Knowledge

<https://arxiv.org/pdf/1803.04042.pdf>

Локальные суррогатные модели Local Surrogate Models

Local Interpretable Model-agnostic Explanations (LIME)

**Для конкретного прогноза чёрного ящика
делаем отклонения – соседей прогноза, собираем датасет
на нём обучаем суррогатную модель**



Ribeiro, Marco Tulio, Singh, Sameer, and Guestrin, Carlos. "why should i trust you?": Explaining the predictions of any classifier. KDD, 2016. <http://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf>

Пример работы LIME на изображениях



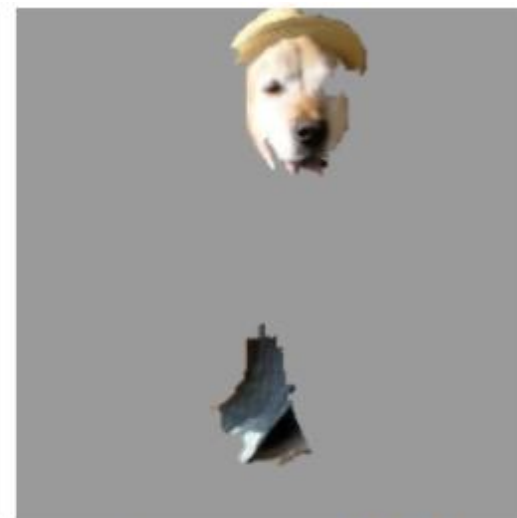
(a) Original Image



(b) Explaining *Electric guitar*



(c) Explaining *Acoustic guitar*



(d) Explaining *Labrador*

НС Inception

Подсвечены супер-пиксели, ответственные за Top-3 класса

Electric Guitar $p=0.32$

Acoustic guitar $p=0.24$

Labrador $p=0.21$

Исследование отдельных блоков модели

За что отвечают отдельные нейроны, каналы, слои

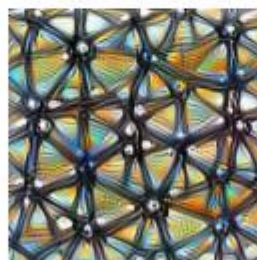
Different **optimization objectives** show what different parts of a network are looking for.

n layer index
 x, y spatial position
 z channel index
 k class index



Neuron

$\text{layer}_n[x, y, z]$



Channel

$\text{layer}_n[:, :, z]$



Layer/DeepDream

$\text{layer}_n[:, :, :]^2$



Class Logits

$\text{pre_softmax}[k]$



Class Probability

$\text{softmax}[k]$

Исследование отдельных блоков модели

Если попробовать сразу красиво не получится... грамотная регуляризация

- **наказывать разницу соседних пикселей**
- **размытие изображения после k шагов**
- **добавление устойчивости к некоторым преобразованиям**
- **априорное распределение на генерируемые изображения**
- **градиентный спуск в другом – декоррелируемом пространстве (в базисе Фурье)**

<https://distill.pub/2017/feature-visualization/>

Подробнее: DL / interesting

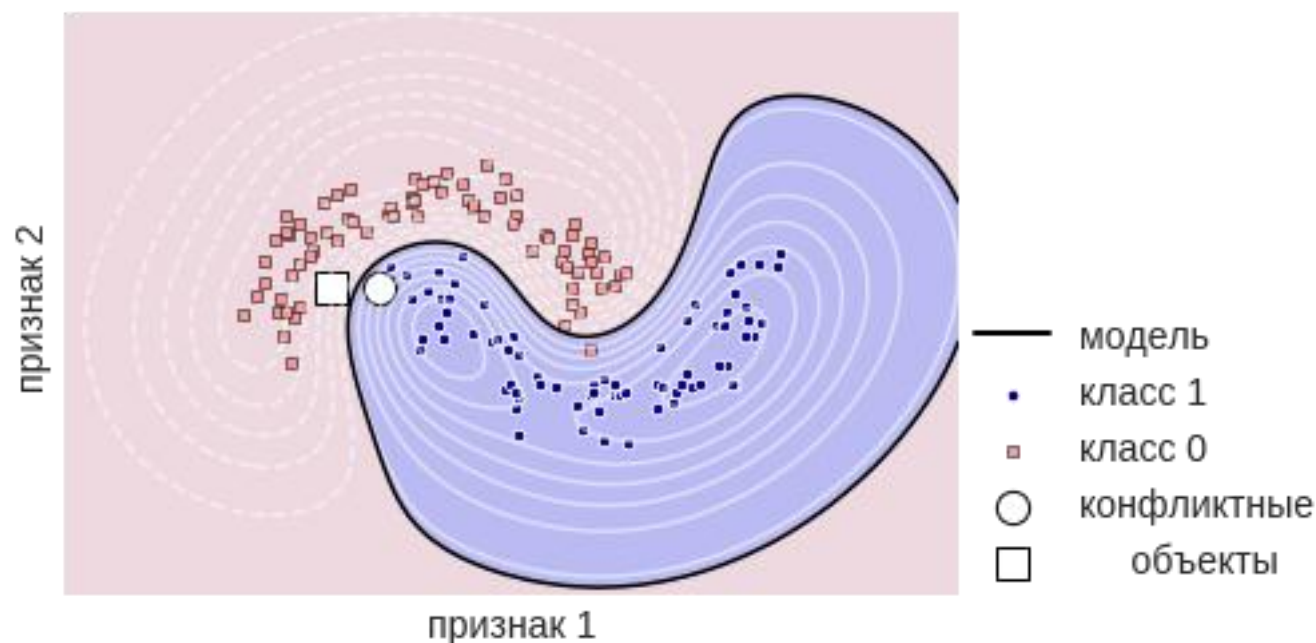
Интерпретация на примерах (Example-based explanations)

- **Конфликтные примеры (Counterfactual explanations)**
- **Adversarial Examples / Attacks (Состязательные примеры / атаки)**
- **Прототипы и критика (Prototypes and Criticisms)**
- **Влиятельные объекты (Influential Instances)**

Конфликтные интерпретации (Counterfactual explanations)

**«если увеличите уровень вашего доход на ... , то получите кредит»
описывают небольшие изменения, которые могут существенно
изменить ответ**

$$\lambda(a(x') - y')^2 + d(x, x') \rightarrow \min_{x'}$$



Конфликтные интерпретации (Counterfactual explanations)

+ понятные интерпретации

+ можно привести пример или указать изменения признаков

+ не нужна выборка

– неединственность решения

– оригинальная реализация имеет ограничения (нет категориальных признаков)

Аналогичное понятие Adversarial Examples, но здесь

- объяснение работы алгоритма**

а там –

- генерация примеров для обмана алгоритма (атаки на сеть)**

Конфликтные интерпретации (Counterfactual explanations) в этой парадигме интересный подход – противоположное понятие какие признаки (якоря / anchors) отвечают за верную классификацию?

т.е. изменение других её не поменяет...



(a) Original image



(b) Anchor for "beagle"



(c) Images where Inception predicts $P(\text{beagle}) > 90\%$



What animal is featured in this picture ?	dog
What floor is featured in this picture?	dog
What toenail is paired in this flowchart ?	dog
What animal is shown on this depiction ?	dog

(d) VQA: Anchor (bold) and samples from $\mathcal{D}(z|A)$

Where is the dog?	on the floor
What color is the wall?	white
When was this picture taken?	during the day
Why is he lifting his paw?	to play

(e) VQA: More example anchors (in bold)

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2018). "Anchors: High-precision model-agnostic explanations." AAAI Conference on Artificial Intelligence
<https://homes.cs.washington.edu/~marcotcr/aaai18.pdf>

неплохой обзор

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR, (1), 1–47
<https://arxiv.org/ftp/arxiv/papers/1711/1711.00399.pdf>

пример с текстами

Martens, D., & Provost, F. (2014). Explaining Data-Driven Document Classifications. MIS Quarterly, 38(1), 73–99. <http://doi.org/10.25300/MISQ/2014/38.1.04>

метод растущих сфер (Growing Spheres)

Laugel, T., Lesot, M.-J., Marsala, C., Renard, X., & Detyniecki, M. (2017). Inverse Classification for Comparison-based Interpretability in Machine Learning
<http://arxiv.org/abs/1712.08443>

Прототипы и критика (Prototypes and Criticisms)

Прототипы – объекты, которые описывают все данные

Критика – примеры данных, которые не согласуются с прототипами



Интерпретация данных, а не модели???

Прототипы и критика (Prototypes and Criticisms)

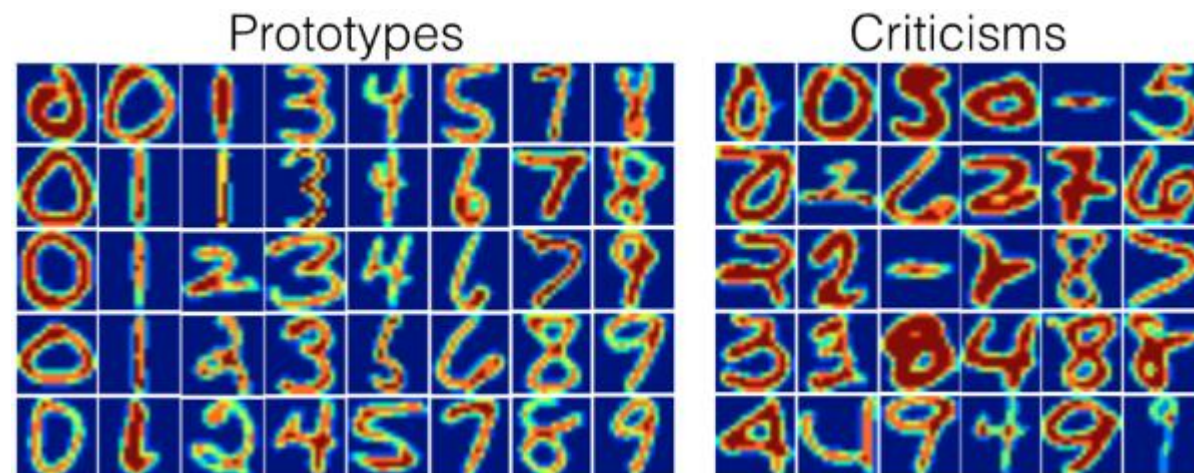
**поиск можно осуществить с помощью кластеризации...
прототипы – центры кластеров**

**Выбрать прототипы так: их распределение ~ распределение данных
Найти критику: там где распределения максимально различаются**

**+ люди лучше понимают задачу,
когда им показывают прототипы и критику**

– надо выбрать число прототипов и критики

**– методы очень специфичные (используют ядерные плотности, не
учитывают качество признаков)**



прототипы



критика



прототипы



критика



Been Kim, Rajiv Khanna, Sanmi Koyejo «Examples are not Enough, Learn to Criticize! Criticism for Interpretability» NIPS 2016

http://people.csail.mit.edu/beenkim/papers/KIM2016NIPS_MMD.pdf

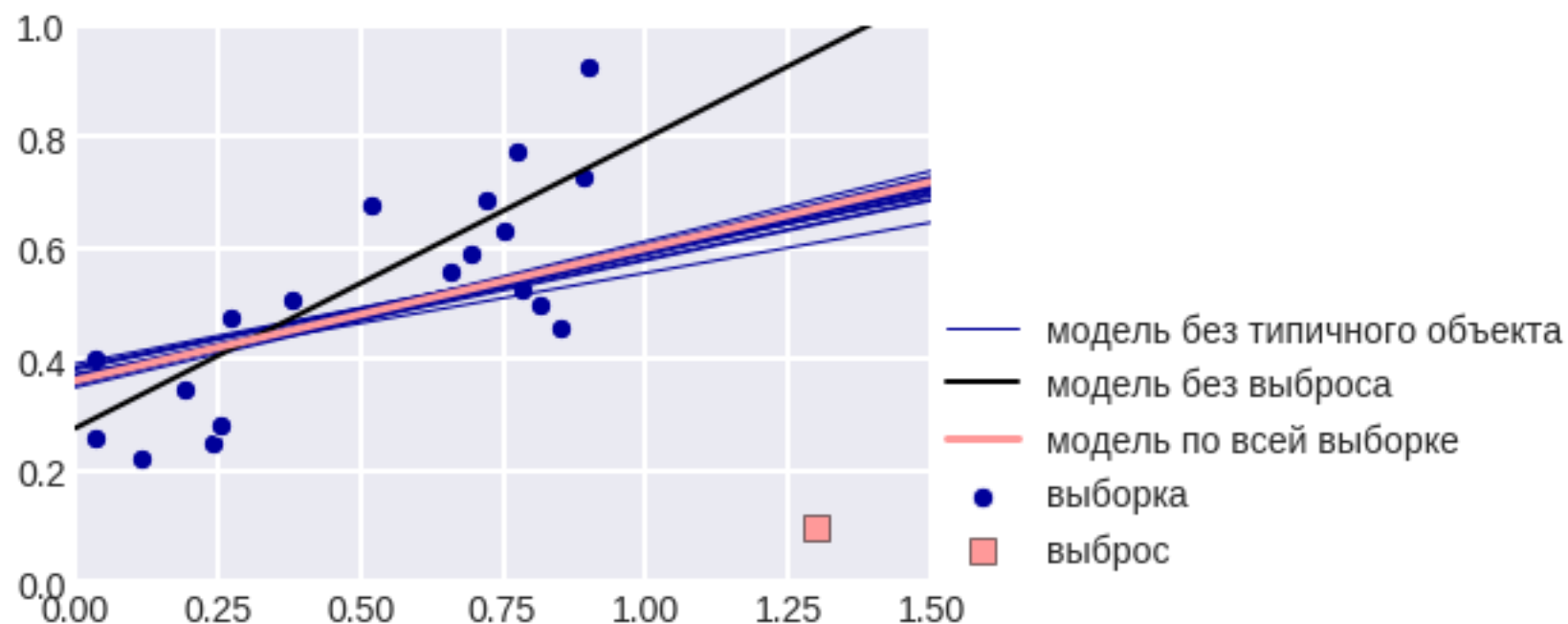
<https://github.com/BeenKim/MMD-critic>

Влиятельные объекты (Influential Instances)

Алгоритм зависит от обучающей выборки

объект / группа объектов может существенно влиять на параметры алгоритма

Часто это аномалии / выбросы



Влиятельные объекты (Influential Instances)

Интерпретация пары: данные – модель

Важно

**не любой влиятельный – аномалия (пример: опорный вектор в SVM)
не любая аномалия – влиятельна (пример: объект с аномальными
значениям признаков, но удовлетворяющий модели)**

Влиятельные объекты (Influential Instances)

Deletion diagnostic

$$\text{influence}(x_i) = \int |a(x | \text{train}) - a(x | \text{train} \setminus \{x_i\})|$$

интегрирование – суммирование по объектам тестовой выборки

– нужно перенастраивать модель

Influence Functions – не нужно перенастраивать

идея: при некоторых предположениях можно вывести аналитически влияние конкретного объекта на функцию ошибки

смотри статью!!!

Koh, P. W., & Liang, P. (2017). Understanding Black-box Predictions via Influence Functions.

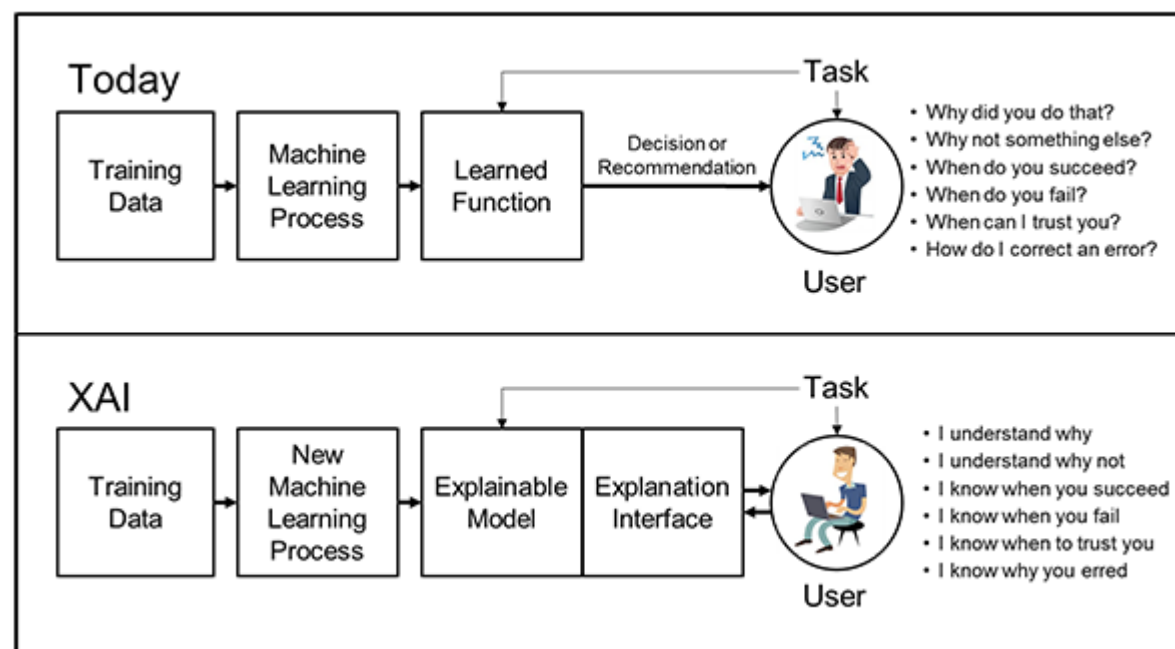
<http://arxiv.org/abs/1703.04730>

<https://github.com/kohpangwei/influence-release>

«Explainable Artificial Intelligence (XAI)»

– название новой программы DARPA

**более понятные модели (с сохранением качества)
дать возможность понимать и «регулировать доверие»**



<https://www.darpa.mil/program/explainable-artificial-intelligence>

Ссылки

А. Дьяконов «Интерпретации чёрных ящиков» // блог

<https://dyakonov.org/2018/08/28/интерпретации-чёрных-ящиков/>

А. Дьяконов «Неправильные интерпретации» // блог

<https://alexanderdyakonov.wordpress.com/2015/07/23/неправильные-интерпретации/>

Отличный обзор

Zachary C. Lipton The Mythos of Model Interpretability 2017

<https://arxiv.org/abs/1606.03490>

Большая книга

Molnar, C. Interpretable Machine Learning 2018

<https://christophm.github.io/interpretable-ml-book/>

Блог наглядного и доступного DL

<https://distill.pub/>

P. Hall и др. «Ideas on interpreting machine learning» // блог

<https://www.oreilly.com/ideas/ideas-on-interpreting-machine-learning>

Библиотеки

SHAP

<https://github.com/slundberg/shap>