Машинное обучение: Объяснительный ИИ (Explainable AI (XAI) or ML)

Объяснительный ИИ (Explainable AI or ML)

Уткин Л.В.

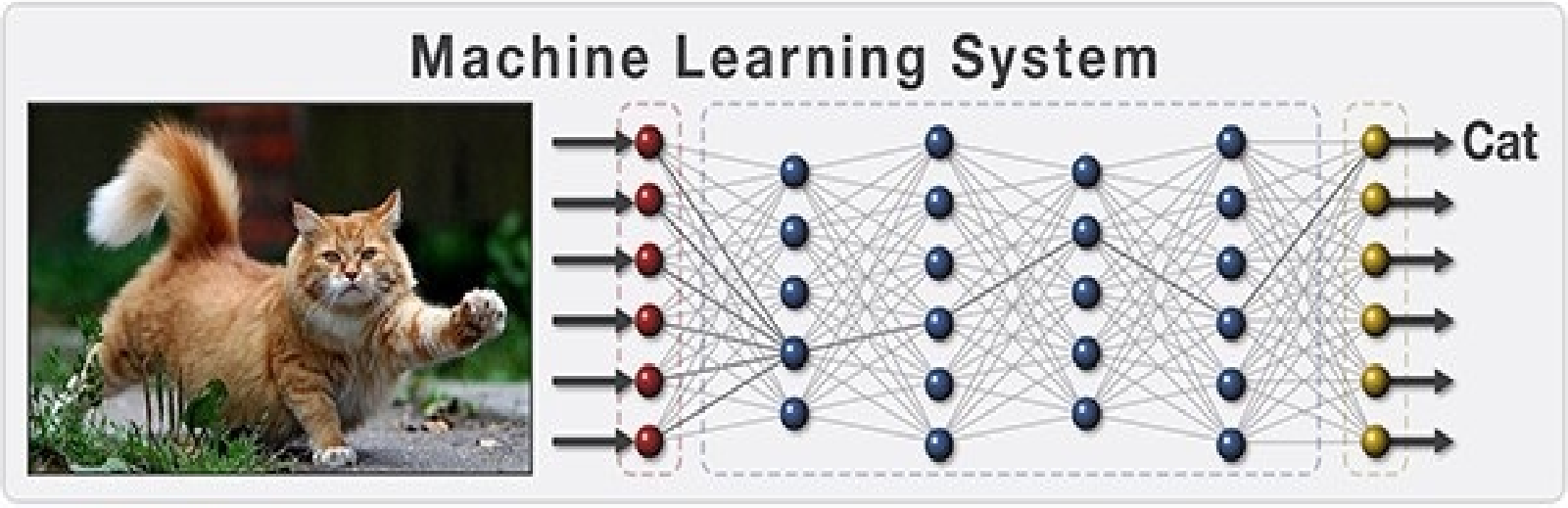
Санкт-Петербургский политехнический университет Петра Великого

Для начала...

"Some things in life are too complicated to explain in any language." ***Haruki Murakami***

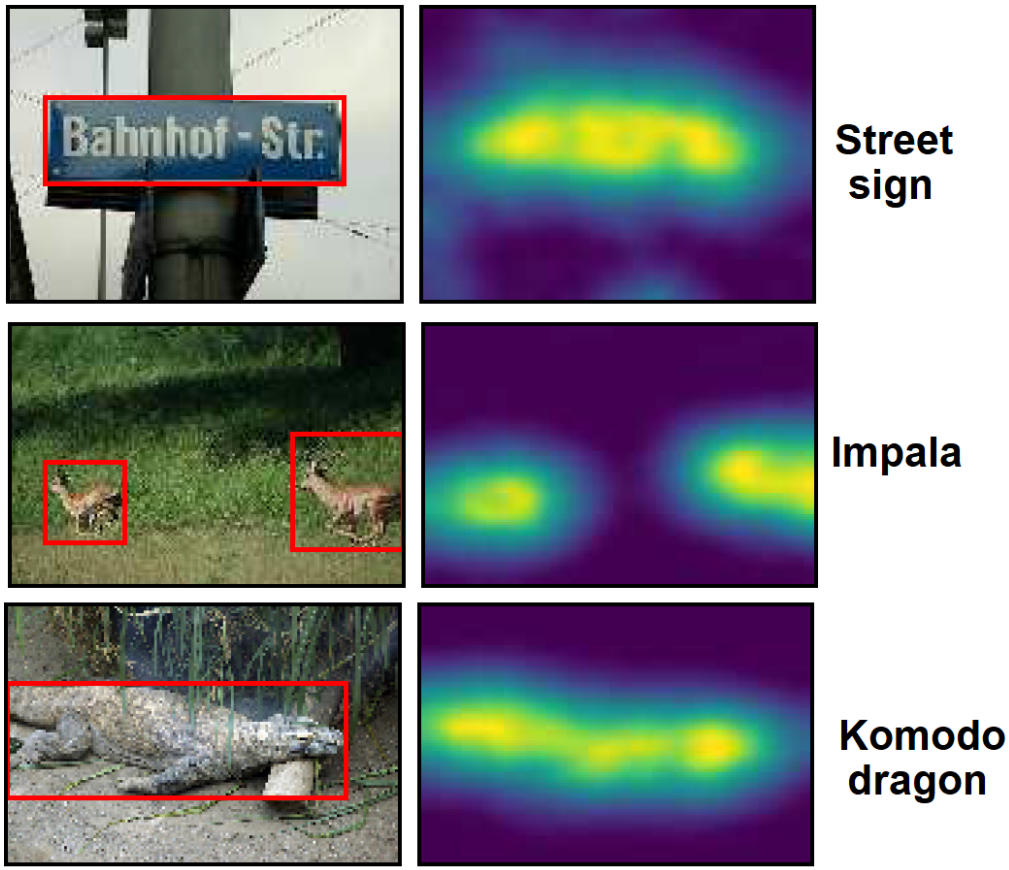
"The solution to explainable AI is more than AI." ***Trevor Darrell***

Что лучше?



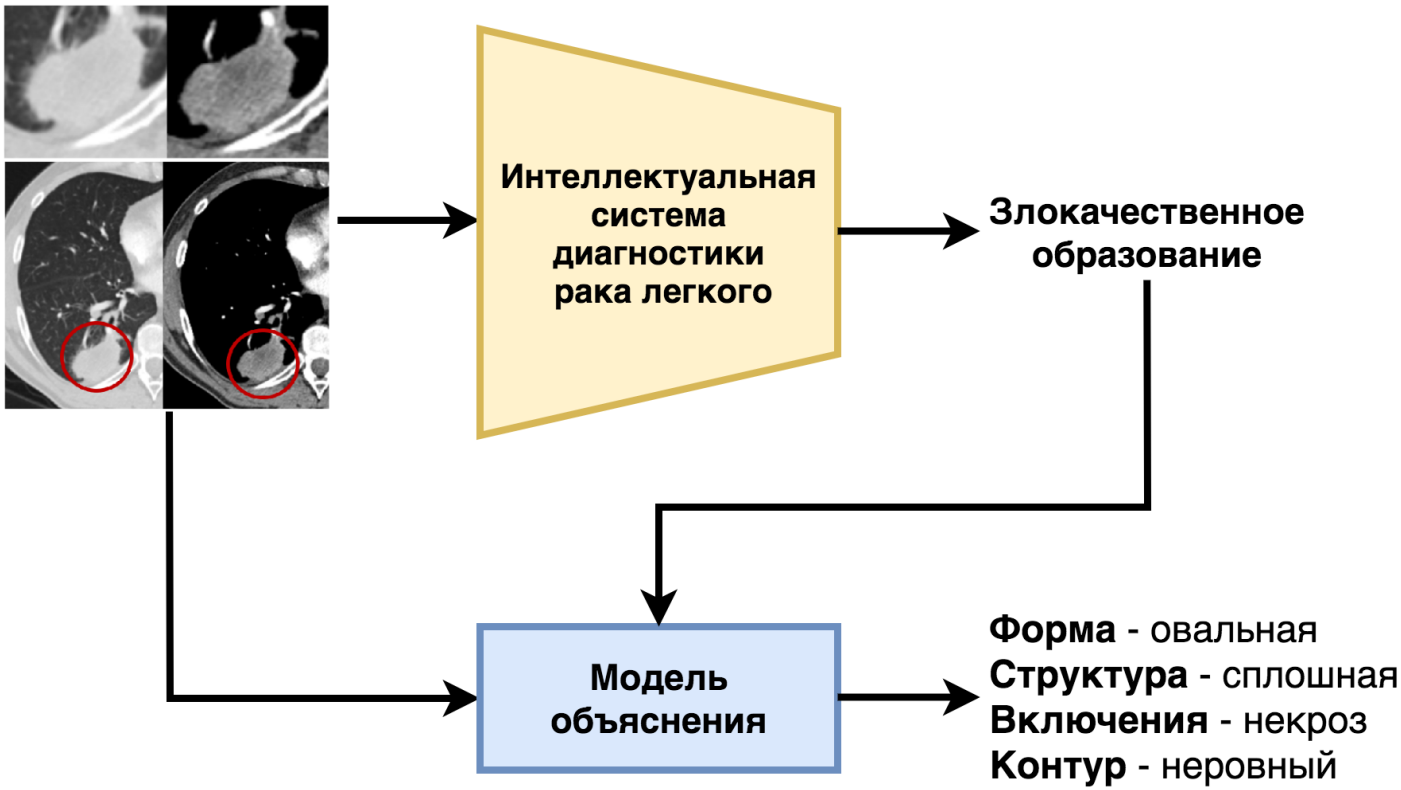
- “модель предсказывает, что это - кот с вероятностью 0.98”
- “модель предсказывает, что это - кот с вероятностью 0.98, так как у него есть шерсть, усы, когти, уши определенной формы”
- как это показать?

Что такое визуальное объяснение?

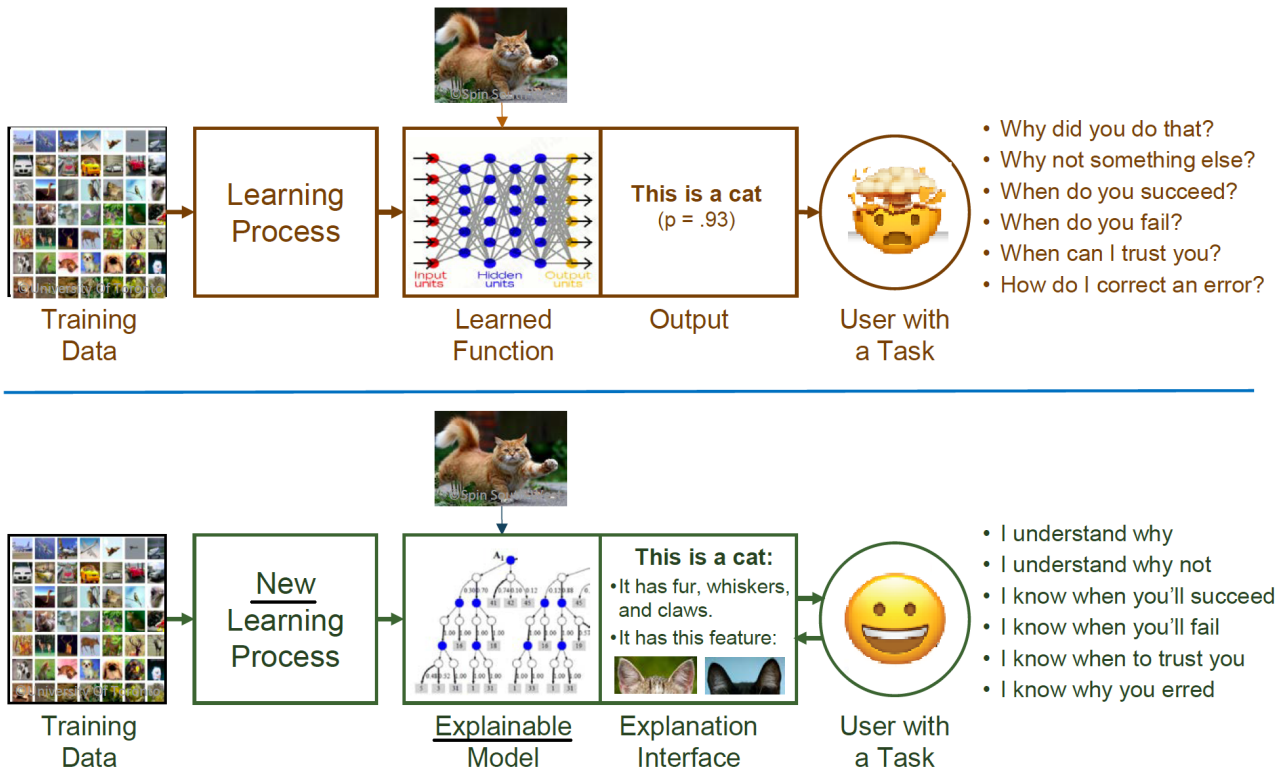


R.C. Fong, A. Vedaldi. Interpretable Explanations of Black Boxes by Meaningful Perturbation, IEEE International Conference on Computer Vision, 2017

Что лучше?


Интеллектуальная система диагностики рака легкого
Злокачественное образование
Модель объяснения
Форма - овальная
Структура - сплошная
Включения - некроз
Контур - неровный

Модели для объяснения (основные) и объяснительные



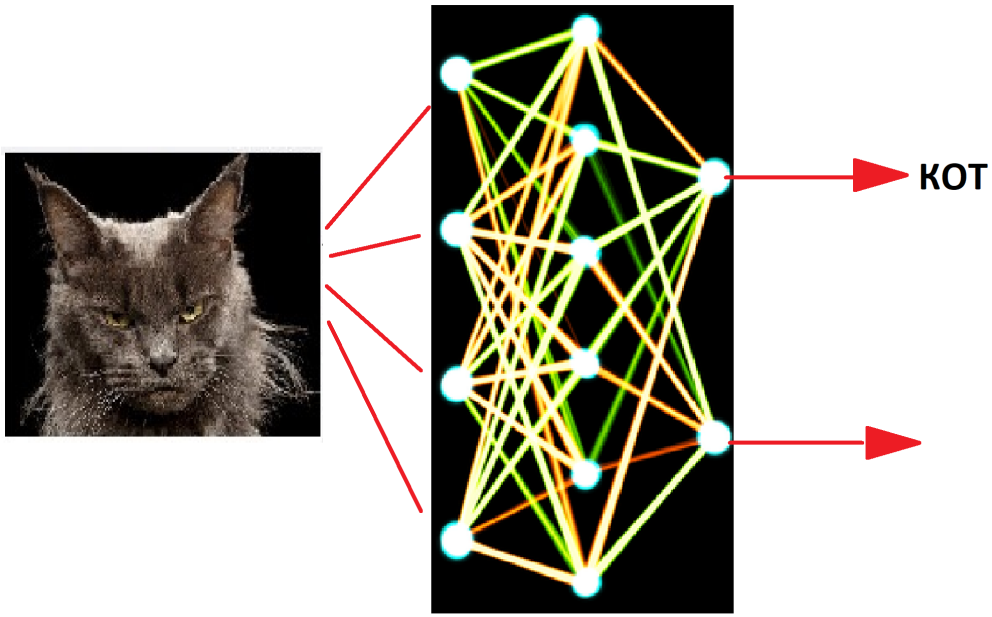
Интерпретация, объяснение, ..., что это и зачем?

- **XAI** (e**X**plainable **A**rtificial **I**ntelligence)
- **Объяснение (explanation)**
- **Интерпретация (interpretation)**
- Прозрачность (transparency)
- Понимание предсказания (understanding)
- Доверие результатам функционирования модели (trust)

Прозрачность, интерпретация, объяснение

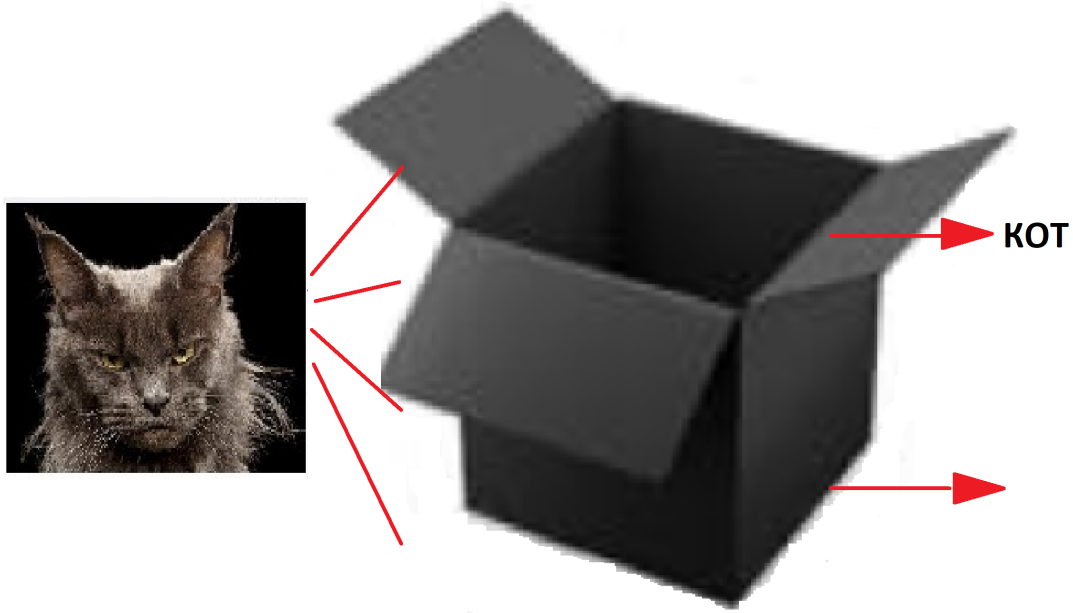
- **Прозрачность (transparency)**: основная модель
  - противоположность “черному ящику”; механизм работы модели известен
  - **данные не используются**
- **Интерпретация (interpretability)**: рассматривает основную модель вместе с данными
  - маски или heatmaps показывают *значимые признаки*
  - **данные всегда используются**
- **Объяснение (explainability)**: рассматривает основную модель, **данные и участие человека**
  - объяснения д.б. интерпретируемы, т.е., давать качественное понимание связи между вх. и вых. данными
  - **цель - объяснить, иструмент - интерпретация**

Стандартная модель МО с точки зрения разработчика



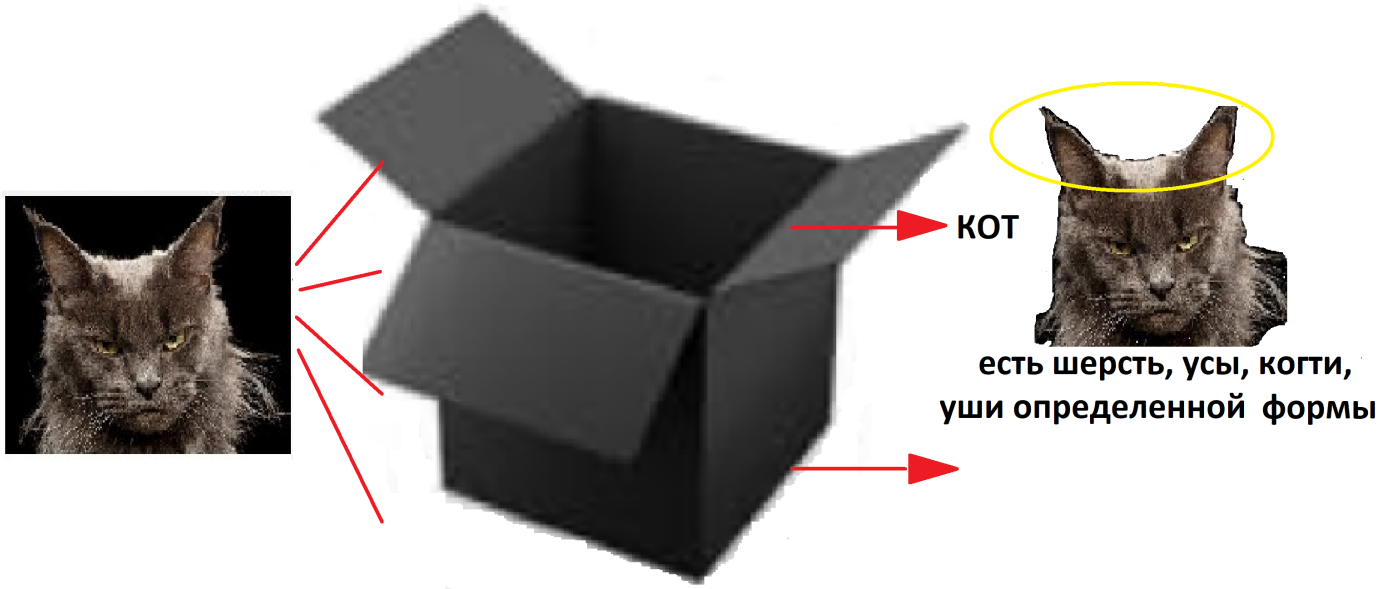
- “модель предсказывает, что это - кот с вероятностью 0.98”

Модель с точки зрения пользователя



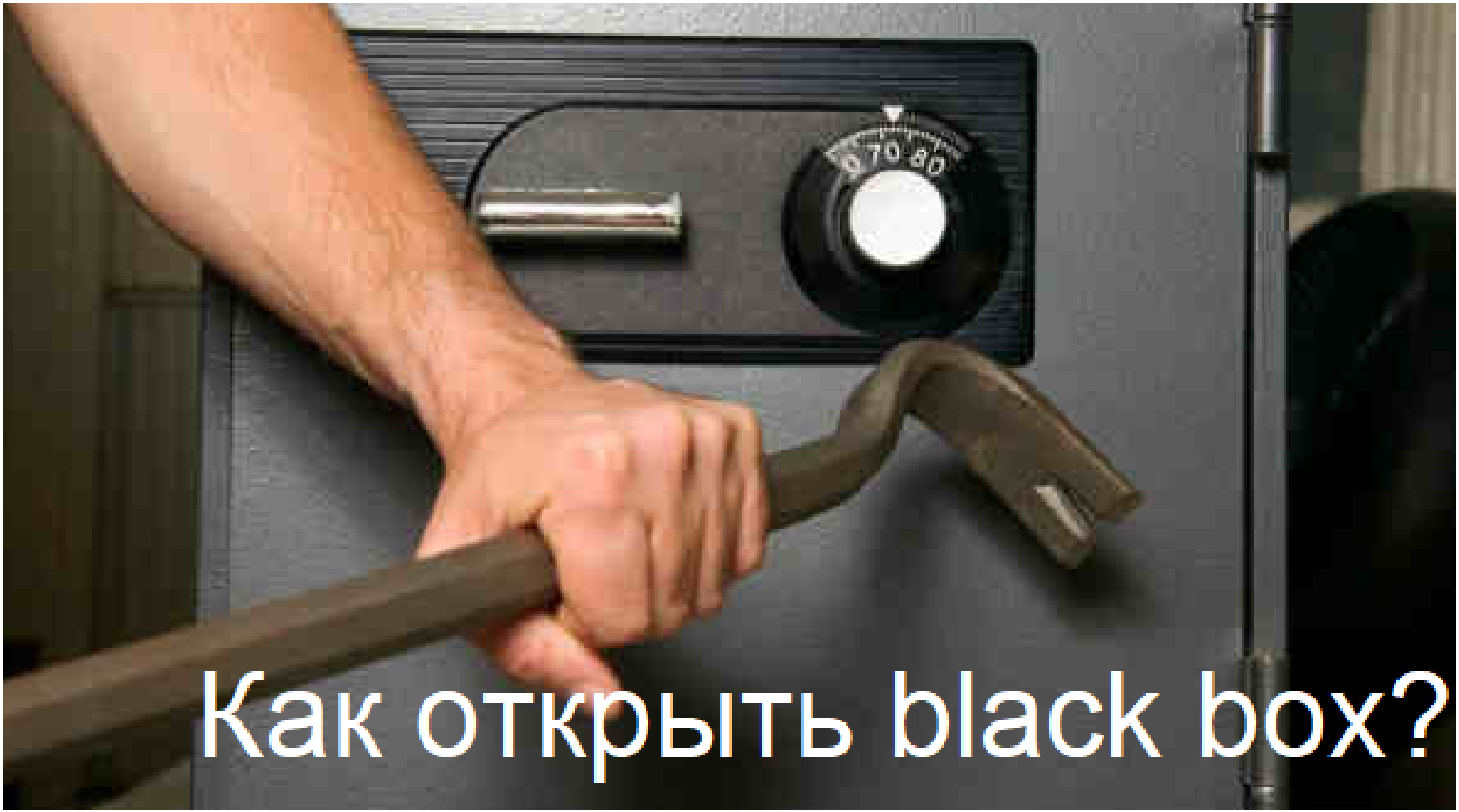
- “модель предсказывает, что это - кот с вероятностью 0.98"

Модель с точки зрения пользователя с объяснением



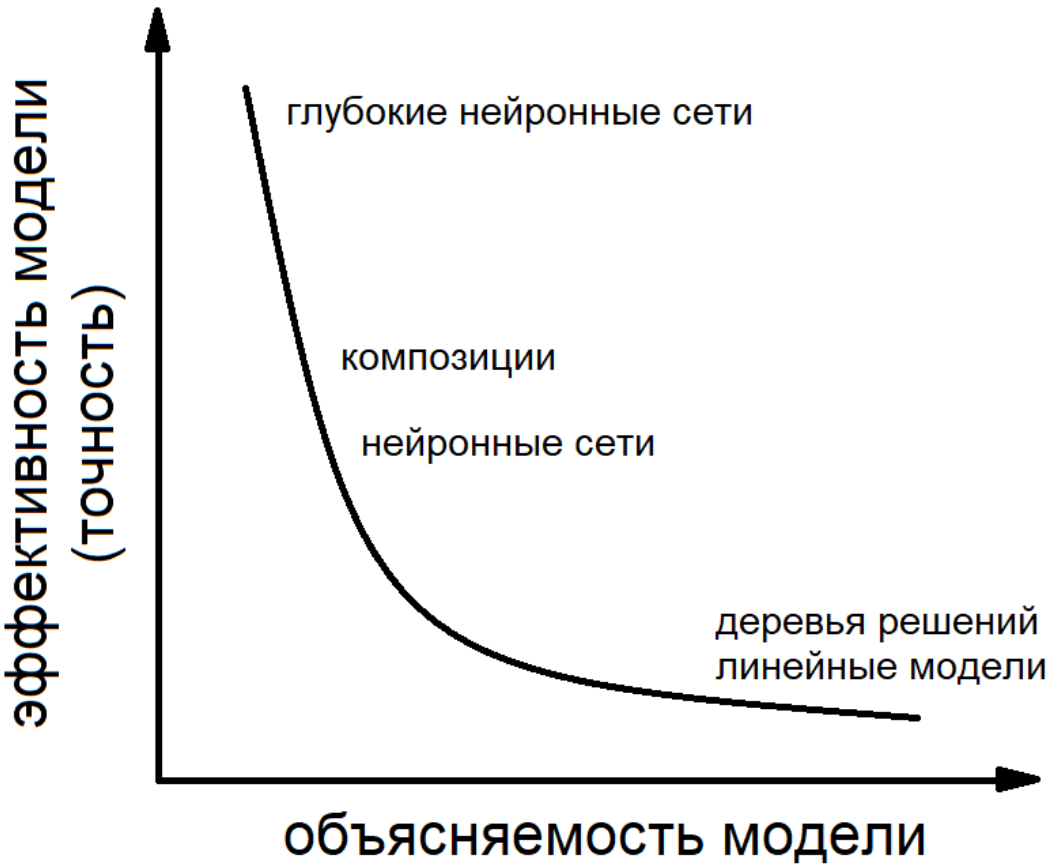
- “модель предсказывает, что это - кот с вероятностью 0.98, так как у него есть шерсть, усы, уши определенной формы”

Как открыть черный ящик и объяснить?



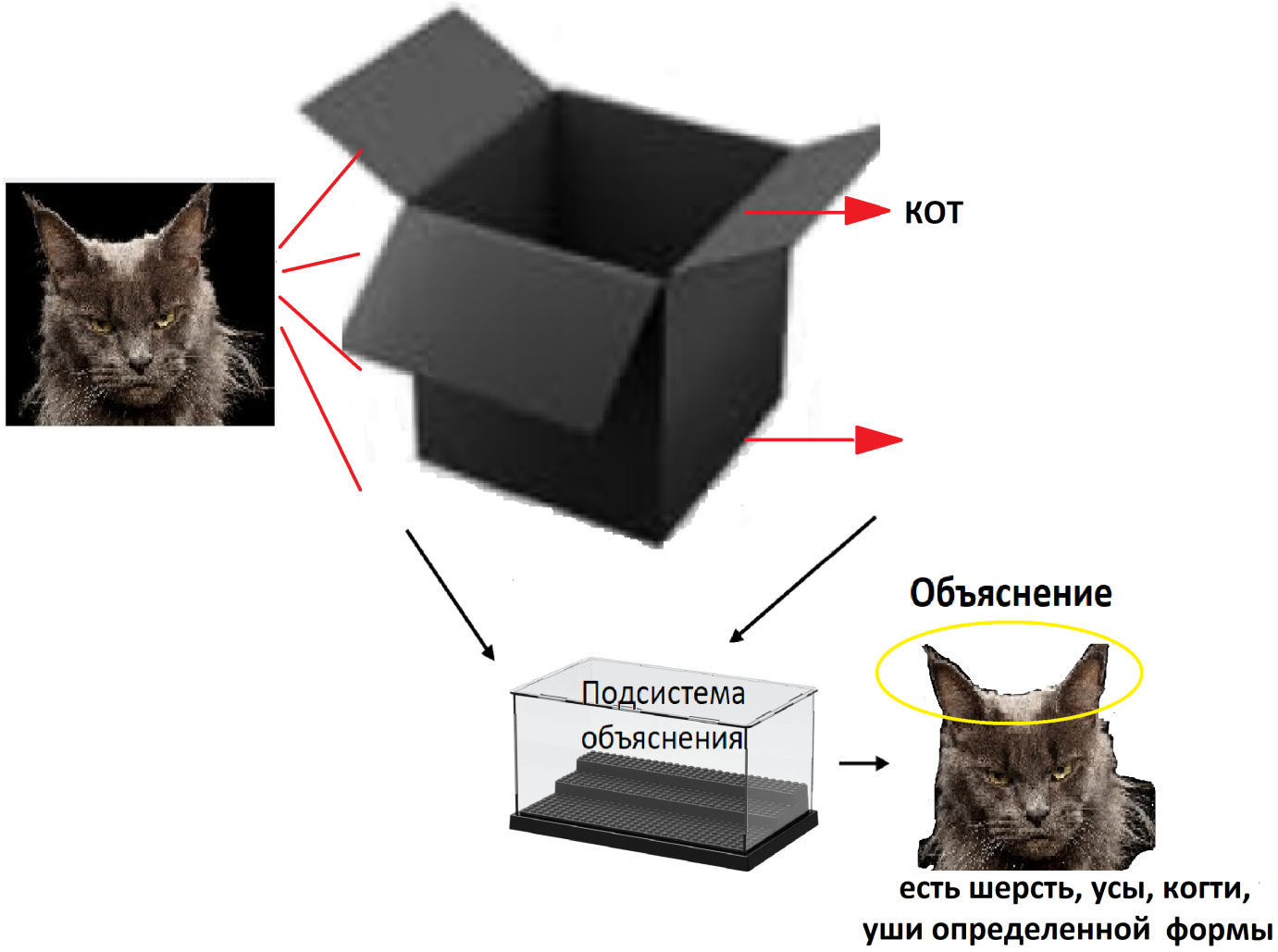
Нужно ли открывать black box?

Нужно, так как эффективные модели обычно не самообъясняемые



Может открыть другой прозрачный ящик?

Модель с точки зрения пользователя с объяснением (другим ящиком)

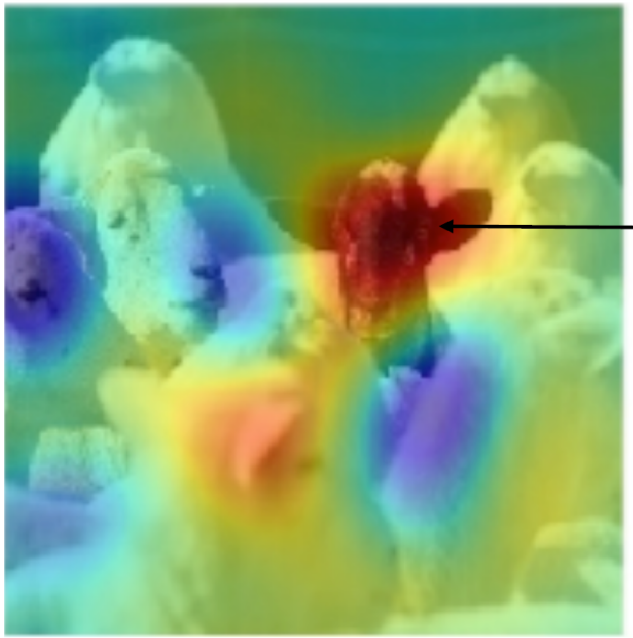


Зачем нужен объяснительный элемент?

Почему классификатор выдает неправильный ответ “корова” для черной овцы?

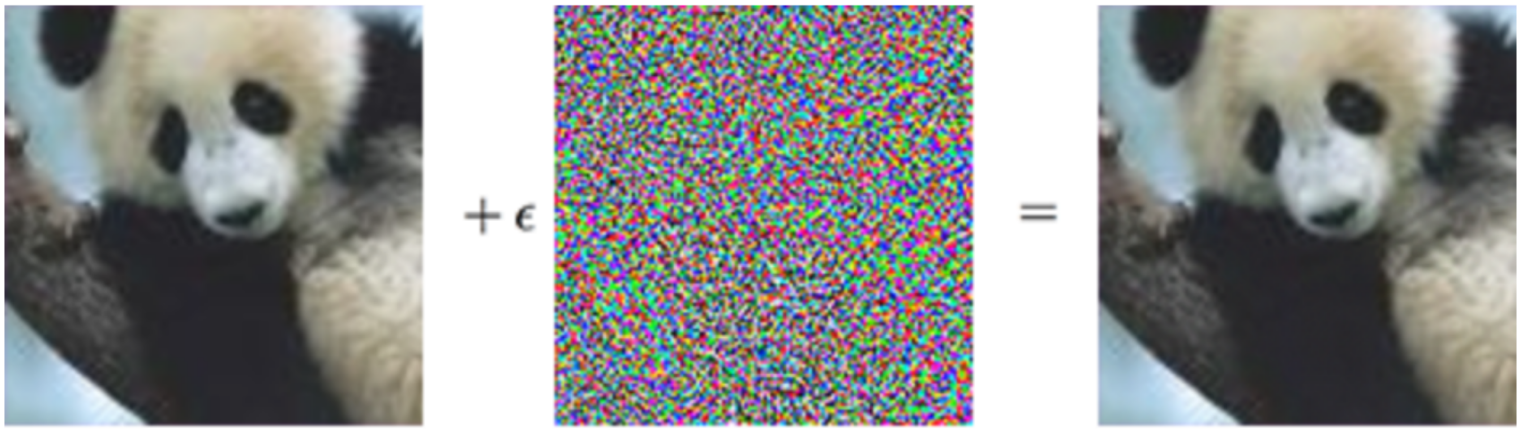
Предсказание: “корова” 81%

Объяснение



Большинство овец - белые, и модель ошибается принимая черную овцу за корову

Зачем нужен объяснительный элемент?



Панда 60% **Гиббон 99%**

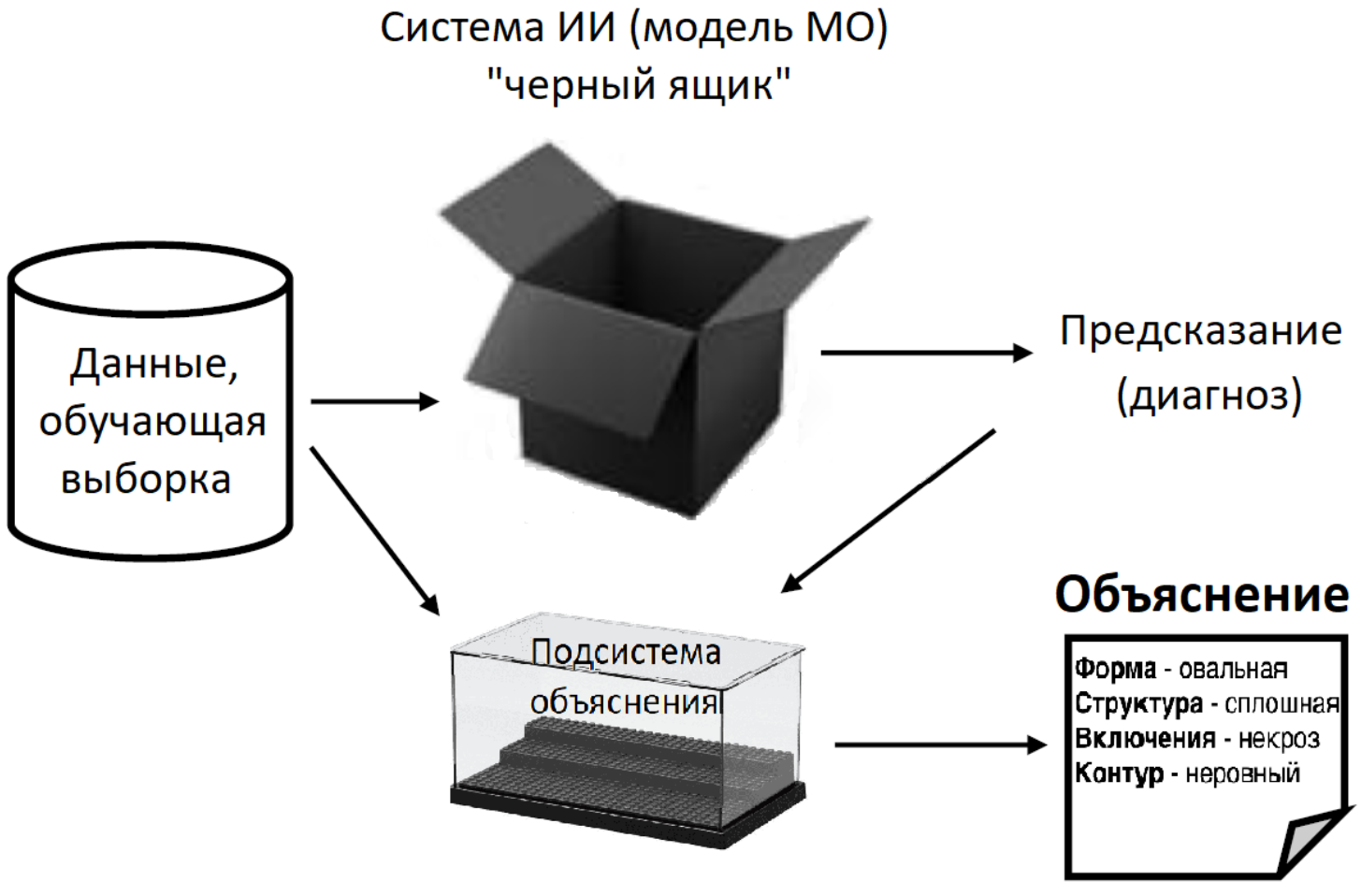
I.J. Goodfellow, J. Shlens, C. Szegedy. Explaining and Harnessing Adversarial Examples // arXiv:1412.6572

Объяснение может помочь лучше понять причины сбоя модели и, в конечном итоге, повысить безопасность системы

Когда или зачем нужен объяснительный элемент?

- Когда прогнозы модели могут иметь далеко идущие последствия, например, рекомендация операции.
- Когда цена ошибки высока, например, неправильная классификация злокачественной опухоли может быть дорогостоящей и опасной.
- Когда выдвигается непонятная гипотеза, например, "Пациенты с пневмонией, страдающие астмой, имели более низкий риск смерти"(Caruana et al. 2015).
- Когда решение подвергается сомнению.
- В любом случае, чтобы люди не могли спрятаться за моделями машинного обучения.

Что нужно в итоге?



Два понятия: интерпретация и объяснение

- **Интерпретация (interpretation)**: рассматривает основную модель вместе с данными
  - маски или heatmaps показывают *значимые признаки*
- **Объяснение (explanation)**: рассматривает основную модель, **данные и участие человека**
  - **наша цель - объяснить, наш иструмент - интерпретация**

Интерпретация и объяснение

Интерпретация

пиксели с номерами от 127 до 194 и от 312 до 391 наиболее значимы для предсказания "КОТ"

Объяснение

уши определенной формы

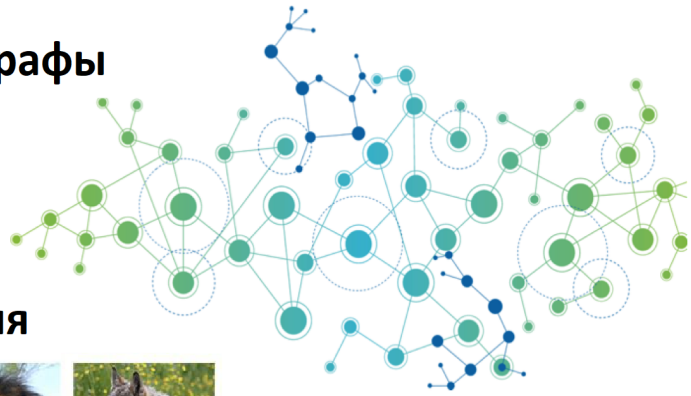
Но и это все условно: интерпретацию и объяснение сложно разделить

Что мы получаем от объяснения? Зависит от данных, которые объясняем

Данные

Текстовые

Графы



Табличные

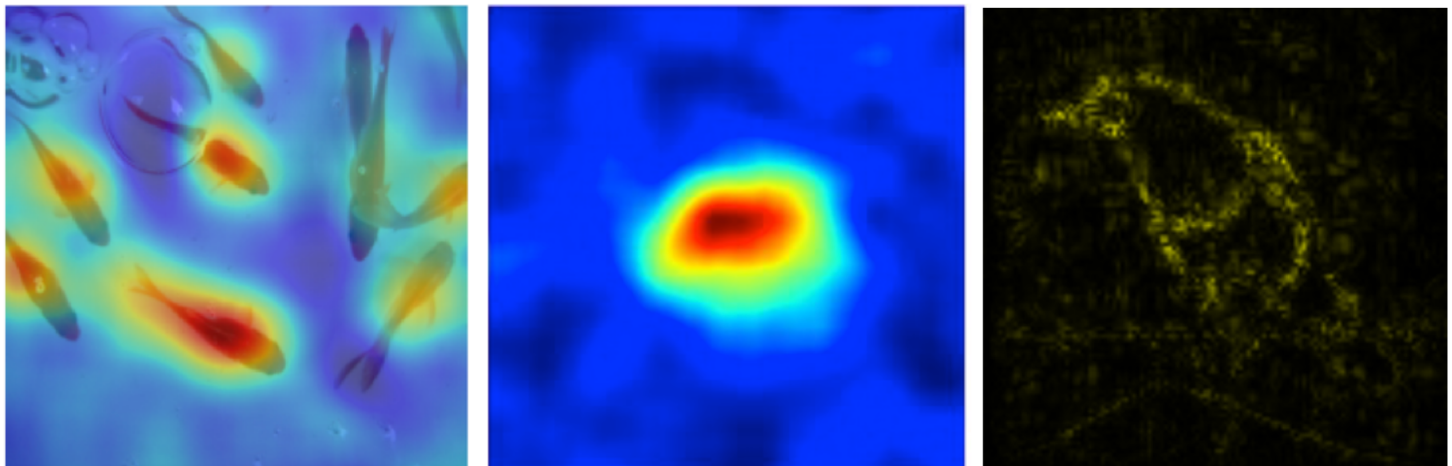
	A	B	C	D
1	Country/City	Bags	Qty	Cost
2	NewYork	2000	200	120000
3	London	1500	200	12000
4	Mexico	1600	200	1213
5	France	1800	100	34523
6	Sydney	1700	150	3454
7	Canada	1650	140	34355
8	Washington	2200	220	5465
9	Glasgow	1900	100	1212
10	Africa	1100	130	3234
11	Auckland	1300	131	1213
12				

Изображения

Способы представления объяснений (данные - изображения)

Изображение

Объяснение (heatmap)



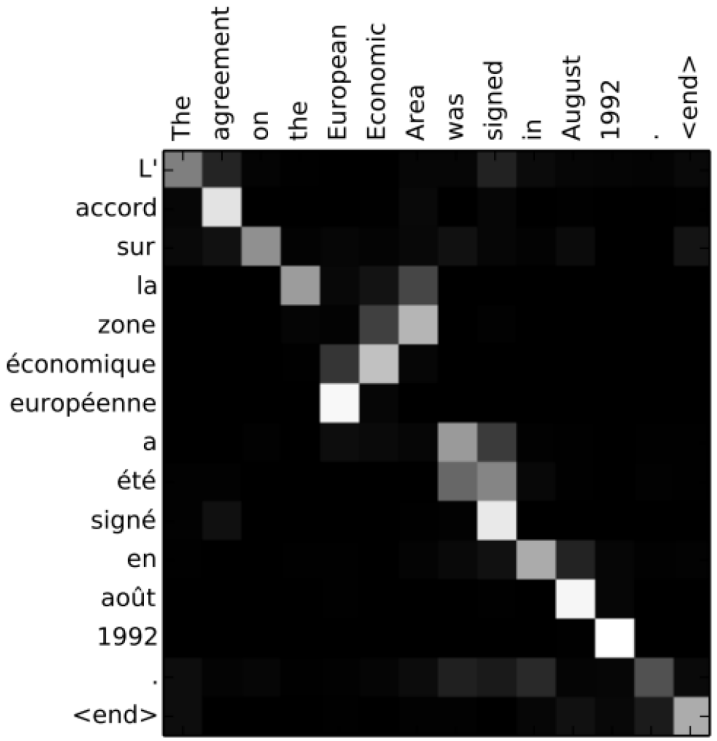
Способы представления объяснений (данные - текст)

Пример из Sentiment Analysis:

Such a great show !

It was a horrible movie

Пример Saliency heatmap из задачи перевода:

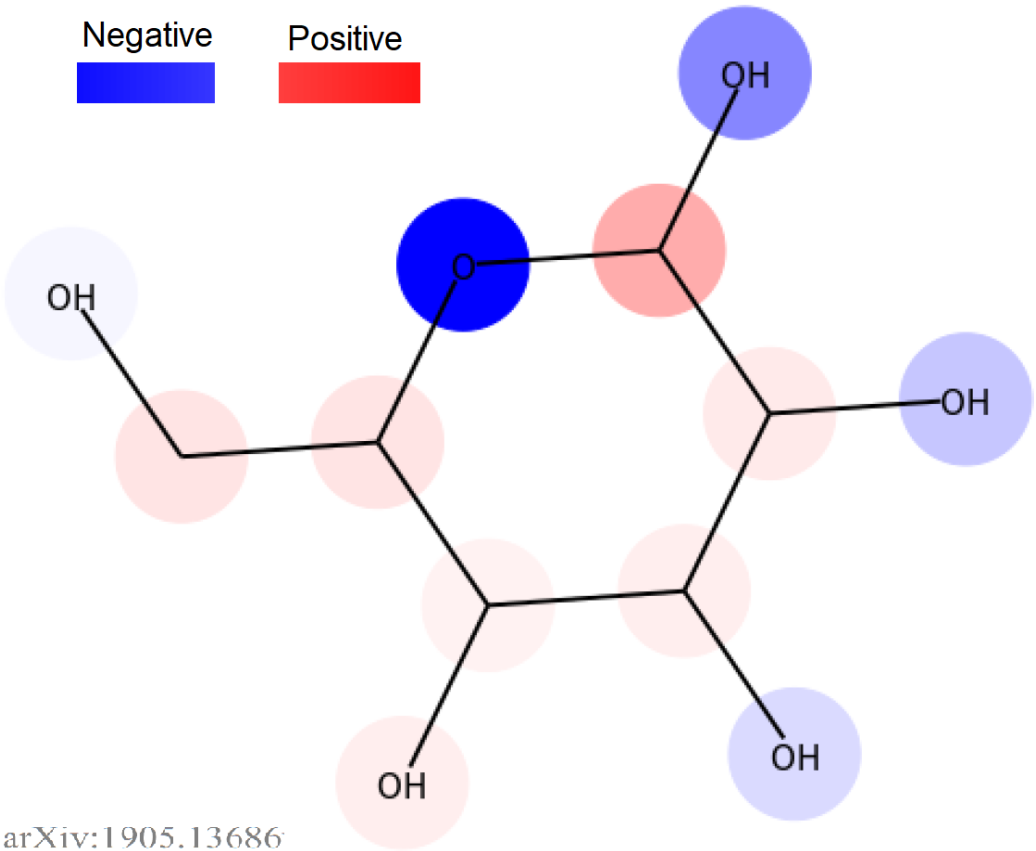


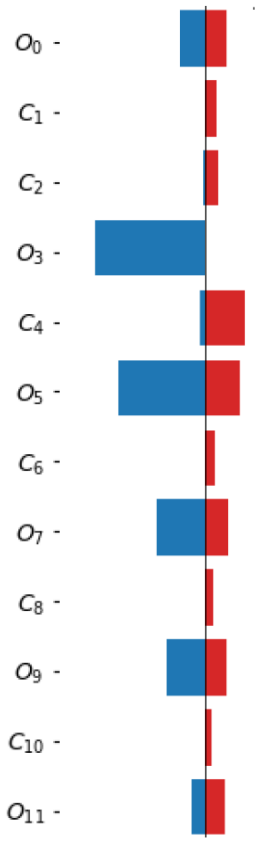
Способы представления объяснений (данные - табличные)

	возраст	дом	доход	образование	кредит
x_1	32	нет	2000	среднее	нет
x_2	54	есть	1200	высшее	да
x_3	73	нет	800	высшее	нет
...	...	...	...	...	...
x_{50}	18	есть	200	среднее	да

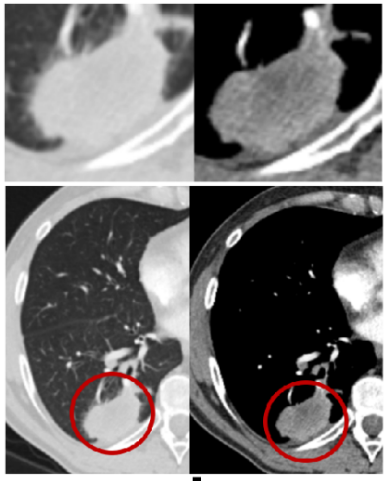
Способы представления объяснений (данные - графы)

Пример предсказания растворимости органических молекул (для глюкозы)





Какое хотелось бы иметь объяснение в идеале?



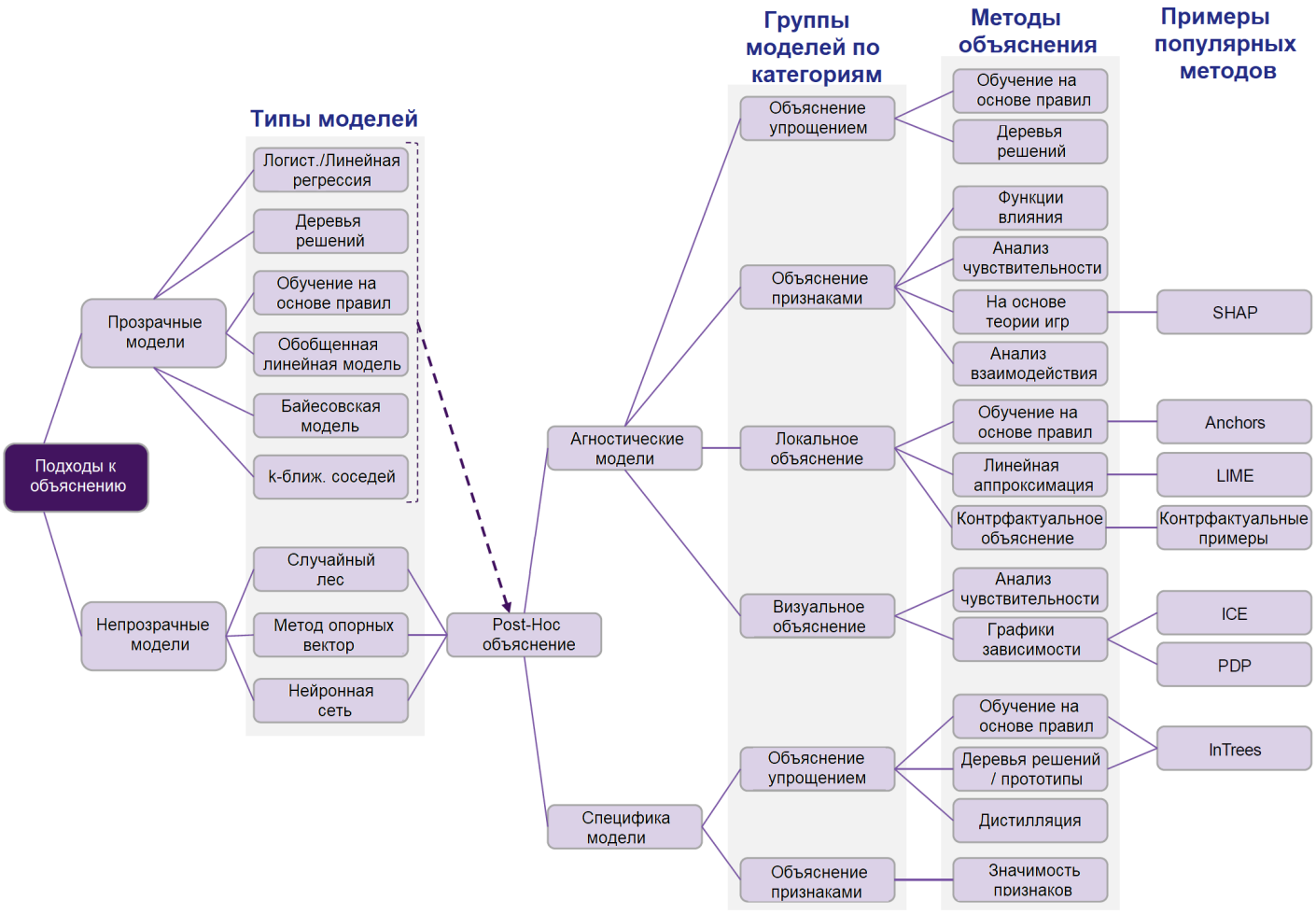
Интеллектуальная система диагностики рака легкого

Злокачественное образование

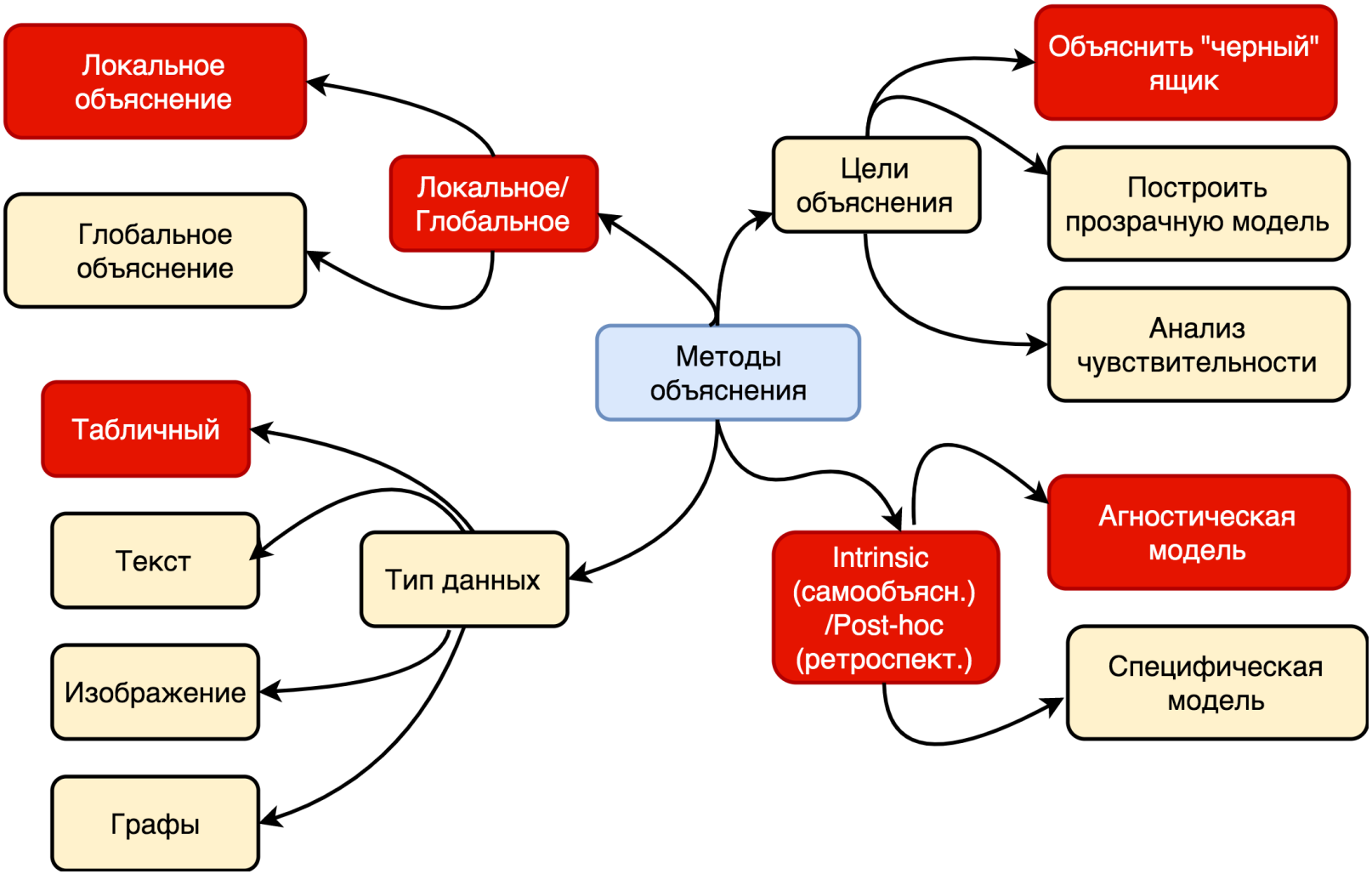
Модель объяснения

Форма - овальная
Структура - сплошная
Включения - некроз
Контур - неровный

Классификация методов объяснения



Классификация моделей для объяснения



Что нас больше интересует?

Локальное объяснение - Post-hoc - “черный ящик” (agnostic)

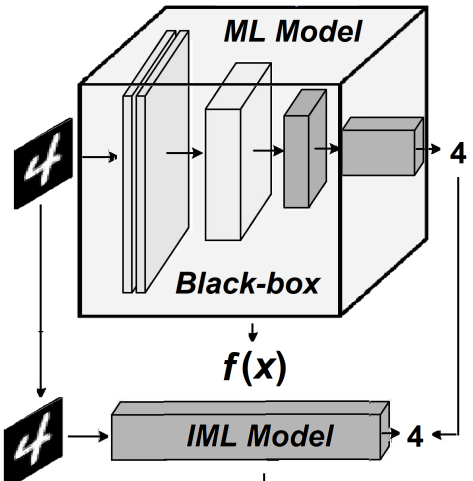
Критерии для методов интерпретации

1. **Post hoc** - объяснение после обучения основной модели
2. **Model-specific** or **model-agnostic**
3. **Local or global**

Post hoc - Model agnostic - Local (модель как "черный ящик")

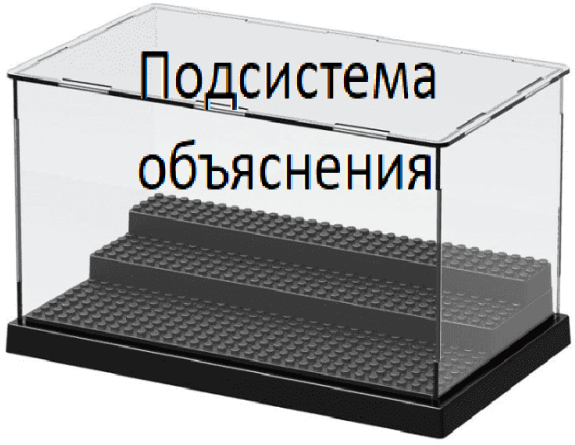
Общая идея локальной интерпретации

Необходимо построить модель (метод) объяснения (объяснитель) для модели МО "черного ящика" (глубокая нейронная сеть, случайный лес, SVM и т.д.), которая ***аппроксимирует*** *основную модель в* ***окрестности*** *объяснимого примера и принадлежит множеству "****простых****" моделей, которые являются* ***самообъясняемыми*** *(линейные модели, деревья решений)*



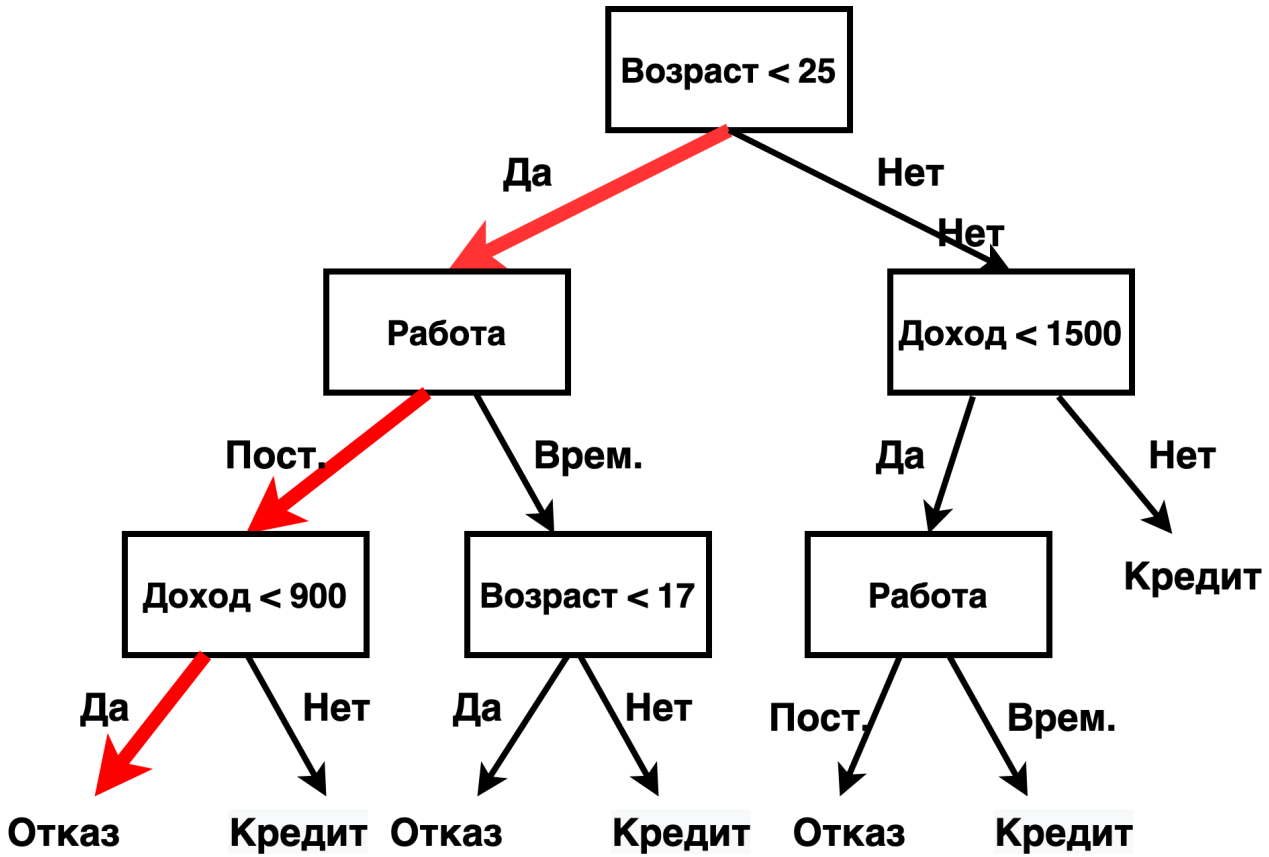
$g(x)$ approximates
$f(x)$ at point $x = x_0$

Объяснительные модели



- **Линейная регрессия**
- Логистическая регрессия
- GLM (Лассо, гребневая регрессия)
- **GAM (Обобщенная аддитивная модель)**
- **Деревья решений**
- **К ближайших соседей**

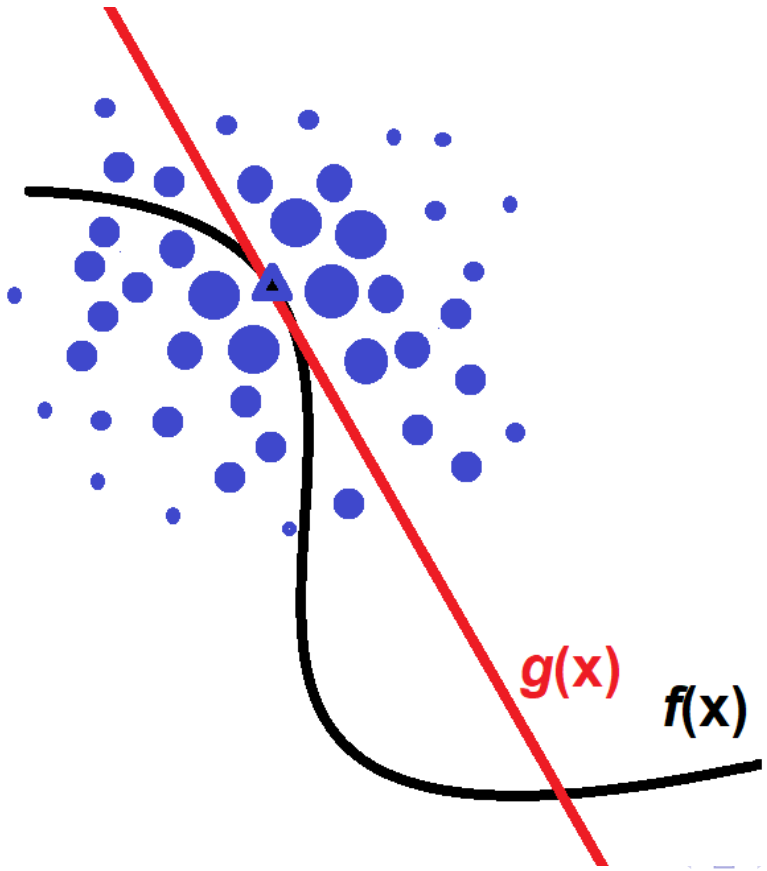
Почему деревья решений?



Почему линейная регрессия?

Линейная регрессия: $g(\mathbf{x}) = a_1 x_1 + a_2 x_2 + \ldots + a_m x_m$

GAM: $g(\mathbf{x}) = g_1(x_1) + g_2(x_2) + \ldots + g_m(x_m)$



Общая модель локальноого объяснения

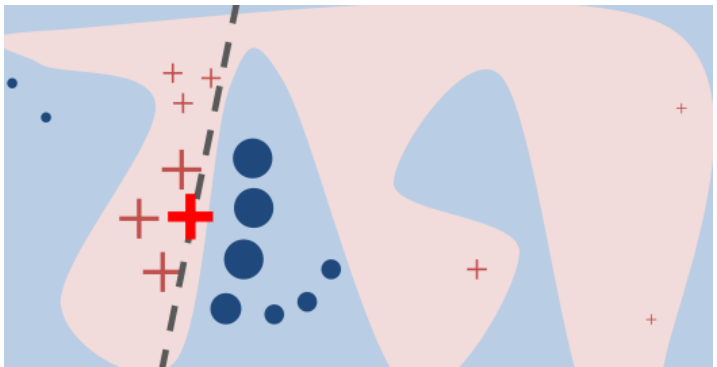
- Основная модель реализует функцию $f : \mathbb{R}^m \to \mathbb{R}^D$, например, в классификации $f(\mathbf{x})$ - вероятность (или индикатор) того, что $\mathbf{x}$ принадлежит определенному классу
- Объяснение - это модель $g \in G$, где G - класс интерпретируемых моделей (линейные модели, GAM, деревья решений)
- Задача оптимизации:

$$\min_{g \in G} \{L(f, g, \theta) + \Omega(g)\}$$

- $L(f, g, \theta)$ - мера того, как неточна g в аппроксимации f
- θ - вектор параметров, $\Omega(g)$ - регуляризатор

Метод LIME (Local Interpretable Model-agnostic Explanations)

1. Основная идея - предположение, что модель **линейная в окрестности анализируемой точки**
2. Вторая идея - возмущение признаков анализируемой точки для генерации новых данных
3. Используя основную модель, находится прогноз $(y = f(x))$ для каждой сгенерированной точки x и образуется **новый датасет**
4. Используя новый датасет, **метод ЛАССО** определяет **значимые признаки**



Метод LIME (3)

- LIME минимизирует функцию

$$\xi = \arg\min_{g \in G} L(f, g, \pi_X) + \Omega(g)$$

- g - объяснительная модель для оригинальной модели f; π_X - веса в виде ядер

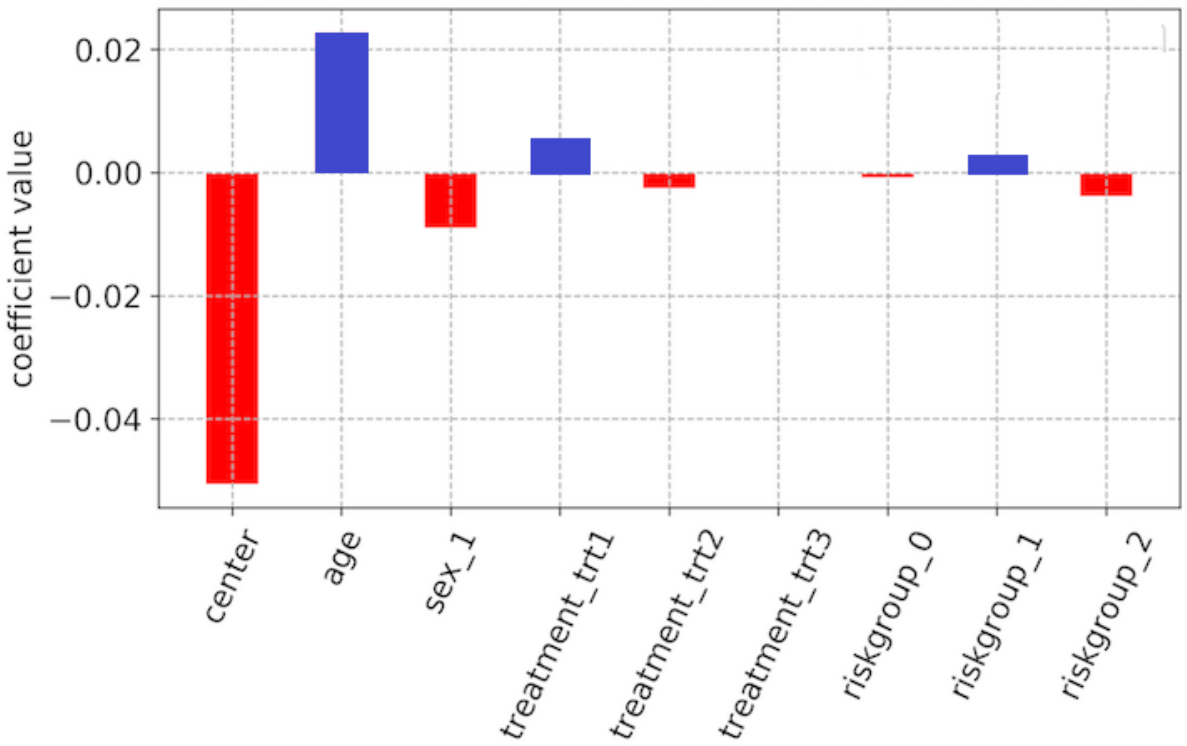
$$g(z) = \phi_0 + \sum_{i=1}^{M} \phi_i z_i$$

LIME (пример)

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

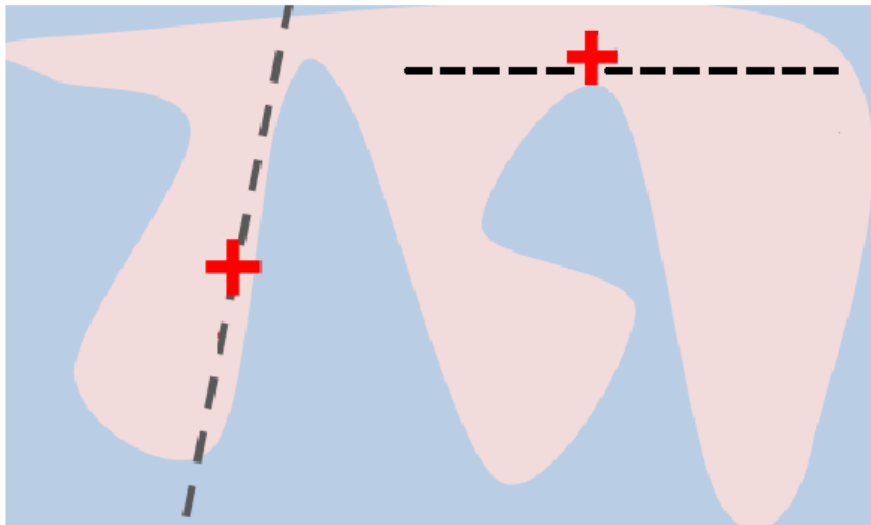
Объяснение вариантов классификации. Три основных прогнозируемых класса: “Electric Guitar” (p = 0.32), “Acoustic guitar” (p = 0.24) и “Labrador” (p = 0.21)

LIME (пример - табличные данные)



LIME (проблемы)

1. Суперпиксели
2. Существенная нелинейность в локальной области



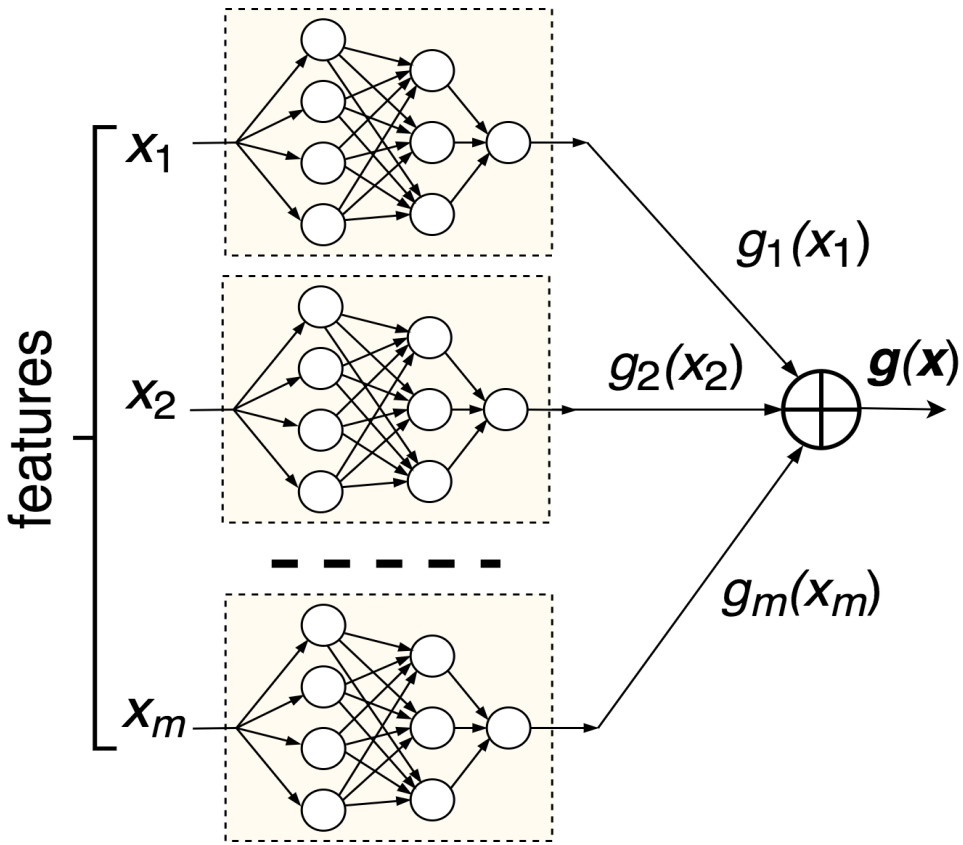
3. Возмущения изображений, текстовые данные

Модификации LIME

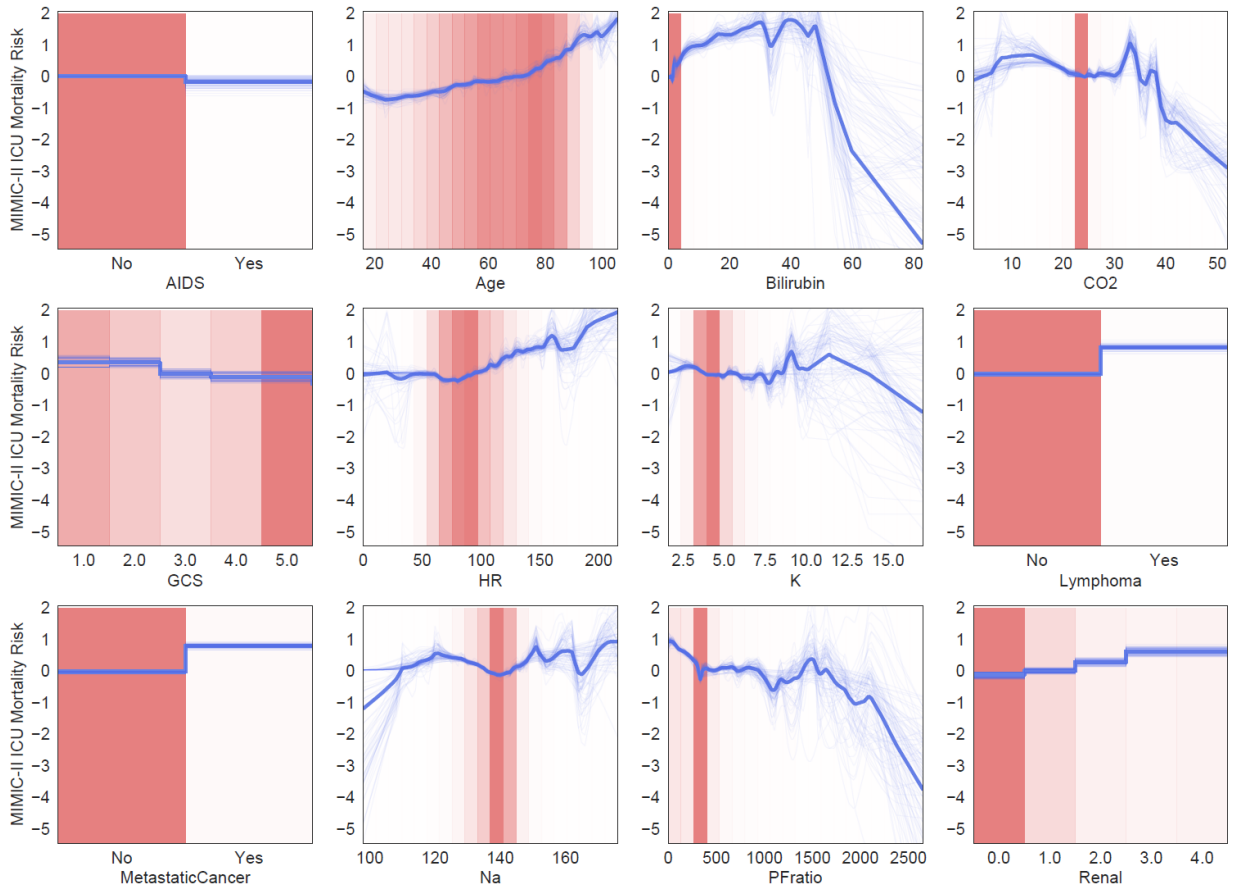
- ALIME (Shankaranarayana and Runje, 2019)
- Anchor LIME (Ribeiro et al., 2018)
- LIME-Aleph (Rabold et al., 2019)
- GraphLIME (Huang et al., 2020)
- SurvLIME (Kovalev et al., 2020)
- SurvLIME-Inf (Utkin et al., 2020)
- и т.д. и т.д.

NAM (neural additive model)

GAM: $g(\mathbf{x}) = g_1(x_1) + g_2(x_2) + \ldots + g_m(x_m)$



NAM: результаты (графики частичной зависимости)



Метод SHAP и теория коалиционных игр

- **SHapley Additive exPlanations (SHAP)**
- Значимости SHAP основаны на числах Шепли, концепции из теории игр
- В теории игр нужны: игра и несколько игроков
- Шепли количественно оценивает вклад каждого игрока в игру

Метод SHAP

- В машинном обучении:
  - «игра» воспроизводит выход (предсказание) модели
  - «игроки» - это признаки (переменные), включенные в модель
- Шепли количественно оценивает вклад каждого игрока в игру
- SHAP количественно оценивает вклад каждого признака в прогноз

Shapley Values

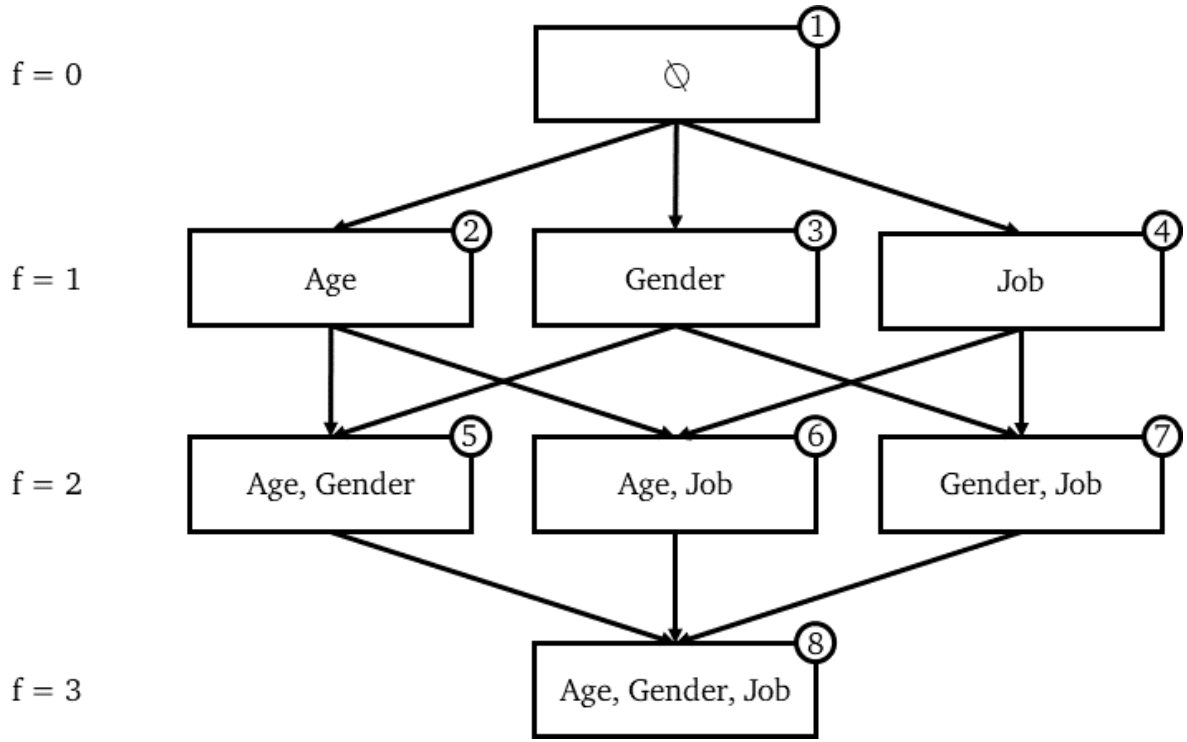
$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}} \left(\mathbf{x}_{S \cup \{i\}} \right) - f_S \left(\mathbf{x}_S \right) \right]$$

- $|F|$ - размер полной коалиции; S - подмножество коалиции, которое не включает игрока i, а $|S|$ - размер S, $S!$ - число перестановок множества S
- В квадратных скобках: «насколько больше выигрыш, когда мы добавляем игрока i к подмножеству S»

SHAP - множество мощности признаков (1)

- Модель машинного обучения: предсказывает доход человека, зная его возраст, пол и работу.
- Числа Шепли основаны на идее, что результат каждой возможной комбинации (или коалиции) игроков должен учитываться для определения важности отдельного игрока.
- В примере это соответствует каждой возможной комбинации f признаков ($f \in \{0, 1, ..., F\}$, $F = 3$).

SHAP - множество мощности признаков (2)



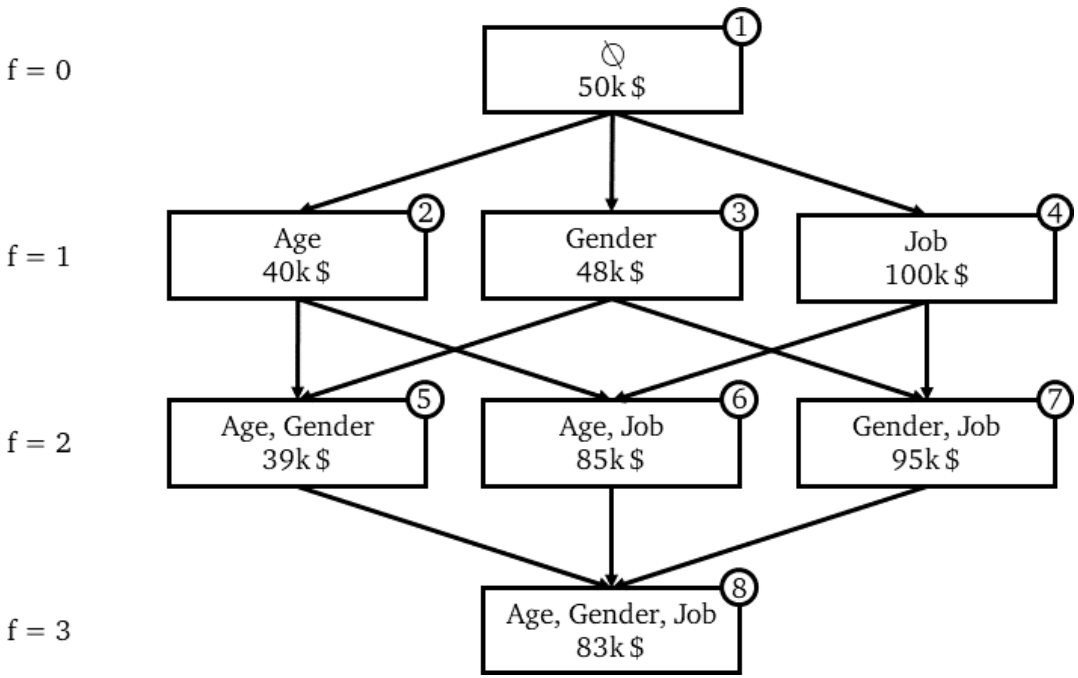
Каждый узел - коалиция признаков, каждое ребро - включение признака, отсутствующего в предыдущей коалиции, 8 коалиций

SHAP - множество мощности признаков (3)

- SHAP требует обучения отдельной модели прогнозирования для каждой отдельной коалиции, то есть 2^F моделей
- Модели полностью эквивалентны друг другу в том, что касается их гиперпараметров и их обучающих данных (которые представляют собой полный набор данных)
- Единственное, что меняется, это набор признаков, включенных в модель.

SHAP - множество мощности признаков (4)

Пусть 8 регрессионных моделей дали 8 прогнозов для x_0



SHAP - маргинальный эффект

- Два узла, соединенные ребром, различаются только одним элементом в том смысле, что нижний имеет точно такие же признаки, что и верхний, плюс дополнительный признак, которого не было у верхнего
- Разрыв между прогнозами двух связанных узлов может быть вменен эффекту этого дополнительного признака
- Это называется «маргинальным вкладом» признака
- И так, каждое ребро - маргинальный вклад, вносимый признаком

SHAP - снова пример

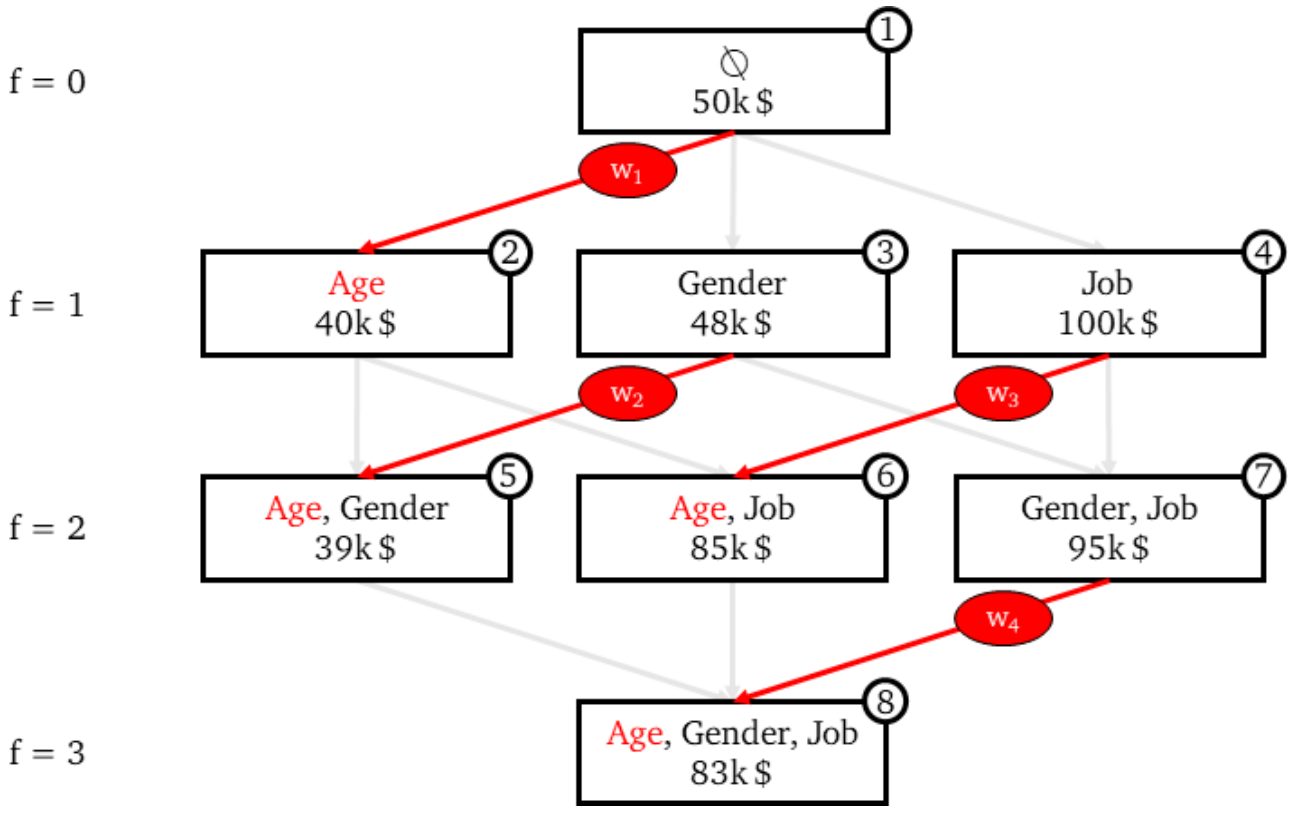
- Представим, что мы находимся в узле 1 (модель без признаков)
- Эта модель предсказывает средний доход от всех обучающих наблюдений: \$50 тыс.
- Если перейдем к узлу 2 (модель только с одним признаком - Age, прогноз для x_0: \$40 тыс.
- Это означает, что знание Age x_0 снизило наш прогноз на \$10 тыс.
- Т.о. маргинальный вклад Age в модель, содержащую только Age в качестве признака: -10k \$

$$\begin{aligned} \mathrm{MC}_{Age,\{Age\}}(x_0) &= \mathrm{Predict}_{\{Age\}}(x_0) - \mathrm{Predict}_{\varnothing}(x_0) \\ &= 40 - 50 = -10 \end{aligned}$$

SHAP - снова пример

- Чтобы получить общий вклад Age на конечную модель (то есть значение SHAP Age для x_0), необходимо учитывать маргинальный вклад Age во всех моделях, где присутствует Age, т.е. рассмотреть все ребра, соединяющие два узла, так что:
  - верхний не содержит Age
  - нижний содержит Age.

SHAP - снова пример



SHAP - снова пример

Все маргинальные вклады затем суммируются с весами:

$$\begin{aligned}\text{SHAP}_{Age}(x_0) = {} & w_1 \times \text{MC}_{Age,\{Age\}}(x_0) \\ & + w_2 \times \text{MC}_{Age,\{Age,Gender\}}(x_0) \\ & + w_3 \times \text{MC}_{Age,\{Age,Job\}}(x_0) \\ & + w_4 \times \text{MC}_{Age,\{Age,Gender,Job\}}(x_0)\end{aligned}$$

где $w_1 + w_2 + w_3 + w_4 = 1$

SHAP - веса ребер

- Сумма весов всех маргинальных вкладов в модели с 1 признаком должна равняться сумме весов всех маргинальных вкладов в модели с двумя признаками и так далее ...
- Т.е. сумма всех весов в том же «ряду» должно равняться сумме всех весов в любом другом «ряду»
- В примере это означает: $w_1 = w_2 + w_3 = w_4$
- Все веса маргинальных вкладов в f-признаковой модели должны быть равны друг другу для каждого f
- Т.е. все ребра одного «ряда» должны быть равны друг другу
- В примере это означает: $w_2 = w_3$
- $w_1 = 1/3, w_2 = 1/6, w_3 = 1/6, w_4 = 1/3$

SHAP - веса ребер

- Спойлер: вес ребра обратно пропорционален общему количеству ребер в одном «ряду».
- Или, что то же самое, вес маргинального вклада в модель f признаков является обратной величиной числа возможных маргинальных вкладов во все модели f признаков.

SHAP - веса ребер

- Каждая модель f признаков имеет f маргинальных вкладов (по одному на каждый признак)
- Достаточно подсчитать число возможных моделей f признаков и умножить его на f.
- Т.о. все сводится к подсчету количества возможных моделей f признаков при заданном f и знании того, что общее количество признаков равно F
- **Это - определение биномиального коэффициента!**

SHAP - веса ребер

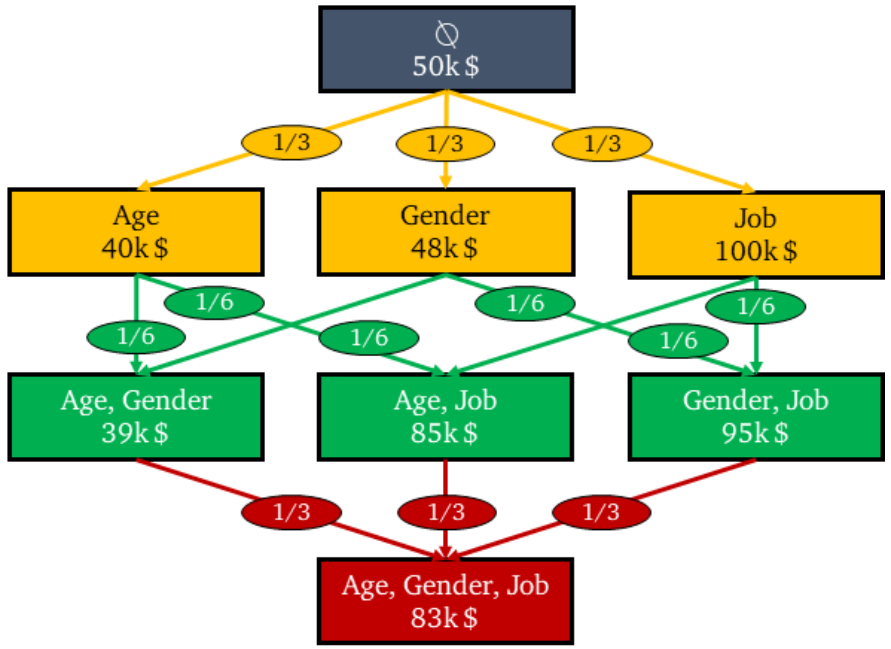
- Итог: количество всех маргинальных вкладов всех моделей f признаков (количество ребер в каждой «строке») - равно:

$$f \times C_F^f$$

- Обратная величина и есть вес маргинального вклада в модель f признаков

SHAP - веса ребер

	N. of Nodes $\binom{F}{f}$	N. of Edges $f \times \binom{F}{f}$
f = 0	1	
f = 1		3
	3	
f = 2		6
	3	
f = 3		3
	1	
Sum	$2^F=8$	$F\times2^{F-1}=12$



SHAP - снова пример

$$
\begin{aligned}
\text{SHAP}_{Age}(x_0) = &\left[1 \cdot C_3^1\right]^{-1} \times \text{MC}_{Age,\{Age\}}(x_0) \\
&+ \left[2 \cdot C_3^2\right]^{-1} \times \text{MC}_{Age,\{Age,Gender\}}(x_0) \\
&+ \left[2 \cdot C_3^2\right]^{-1} \times \text{MC}_{Age,\{Age,Job\}}(x_0) \\
&+ \left[3 \cdot C_3^3\right]^{-1} \times \text{MC}_{Age,\{Age,Gender,Job\}}(x_0) \\
&= \frac{1}{3}(-10) + \frac{1}{6}(-9) + \frac{1}{6}(-15) + \frac{1}{3}(-12) = -11.33\$
\end{aligned}
$$

Shapley Values (1)

$$\text{SHAP}_{feature}(x) = \sum_{set:feature \in set} \left[|set| \times C_F^{|set|} \right]^{-1} \times \left[\text{Predict}_{set}(x) - \text{Predict}_{set \setminus feature}(x) \right]$$

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}} \left(x_{S \cup \{i\}} \right) - f_S \left(x_S \right) \right]$$

- $|F|$ - размер полной коалиции; S - подмножество коалиции, которое не включает игрока i, а $|S|$ - размер S, $S!$ - число перестановок множества S
- В квадратных скобках: «насколько больше выигрыш, когда мы добавляем игрока i к подмножеству S»

Shapley Values (2)

- А как теперь с признаками?
- Вклад i-го признака:

$$\phi_i = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} \left[f_x\left(z'\right) - f_x\left(z' \backslash i\right) \right]$$

- M - общее число признаков; z' - подмножество признаков, которое является объяснением
- Оцениваем значение модели с и без i-го признака ($f_x\left(z'\right)$ и $f_x\left(z' \backslash i\right)$)

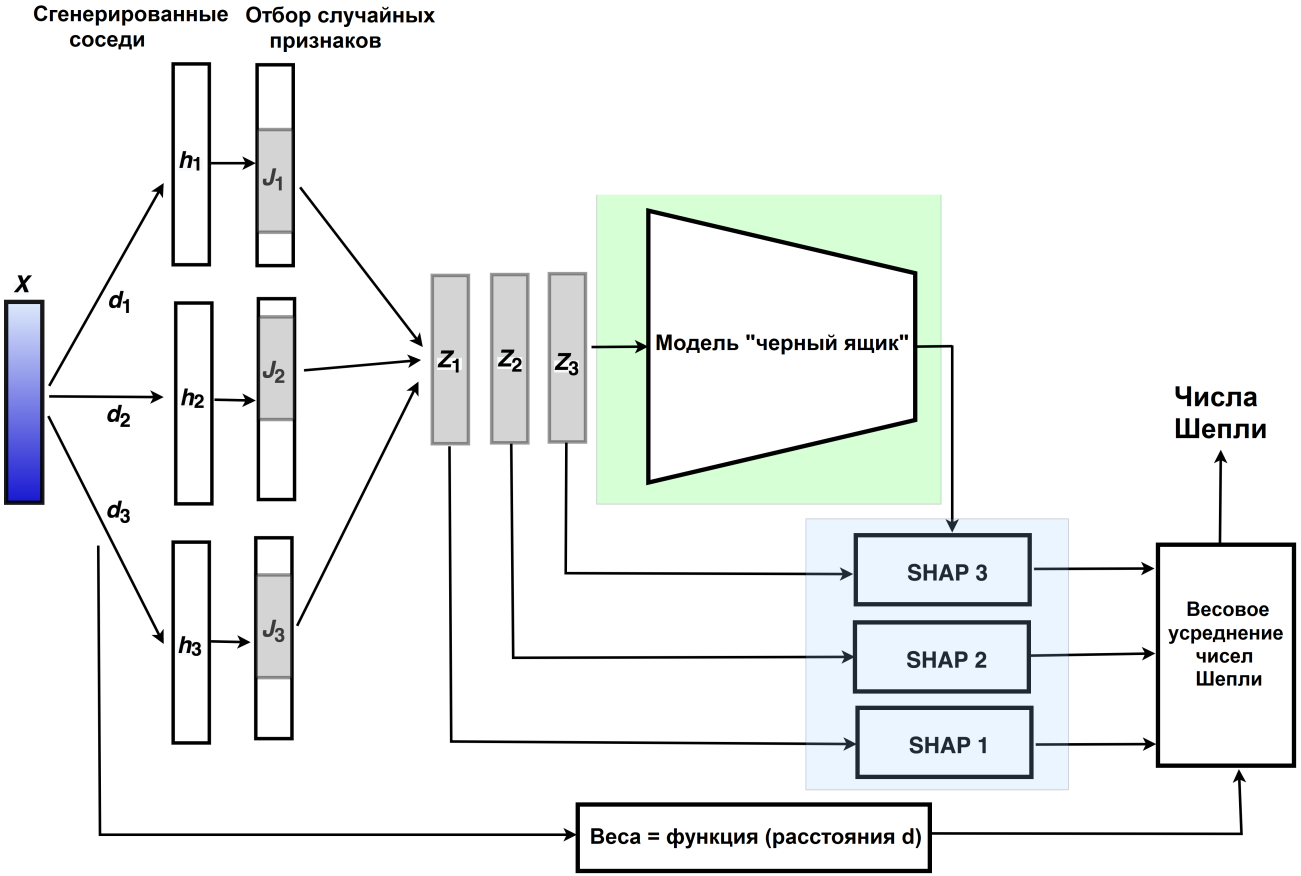
Shapley Values - снова пример

- $\text{SHAP}_{Age}(x_0) = -11.33$, $\text{SHAP}_{Gender}(x_0) = -2.33$, $\text{SHAP}_{Job}(x_0) = 46.66$
- Сумма = + \$33 тыс. В точности равно разнице между выходом всей модели (\$83 тыс.) и выходом пустой модели без признаков (\$50 тыс.)
- **Фундаментальная характеристика чисел SHAP**: сумма чисел SHAP каждого признака наблюдения дает разность между прогнозом модели и нулевой моделью (**SHapley Additive exPlanations**)

SHAP (проблемы)

1. Каким образом заменять признаки вне подмножеств S, т.е. как заполнить данные до полной коалиции?
2. Время вычислений: для каждого числа Шепли необходимо перебрать $2^{|F|}$ вариантов коалиций (подмножеств признаков) и “прогнать” каждый вариант через “черный ящик”. Если признаков больше 30, то задача практически не решается.

Random SHAPs (случайные SHAPы)



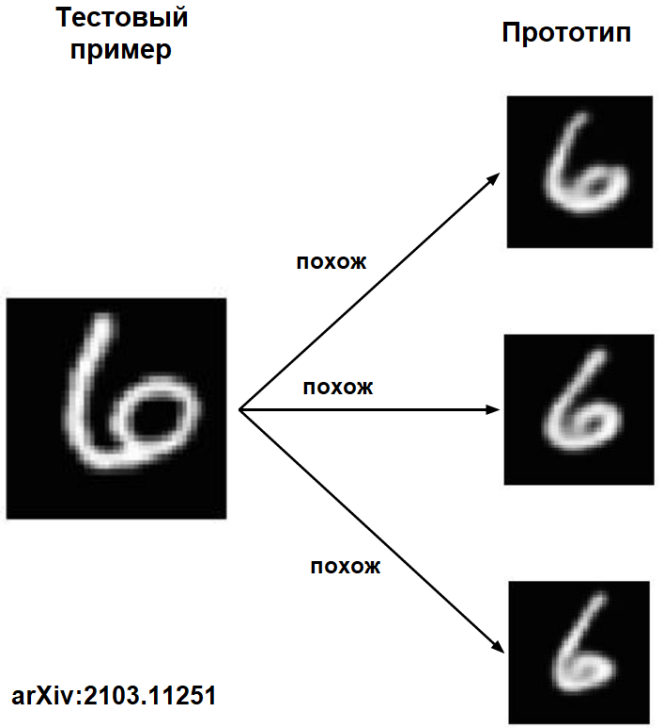
Utkin L.V., Konstantinov A.V. Ensembles of Random SHAPs // arXiv:2103.03302, Mar., 2021

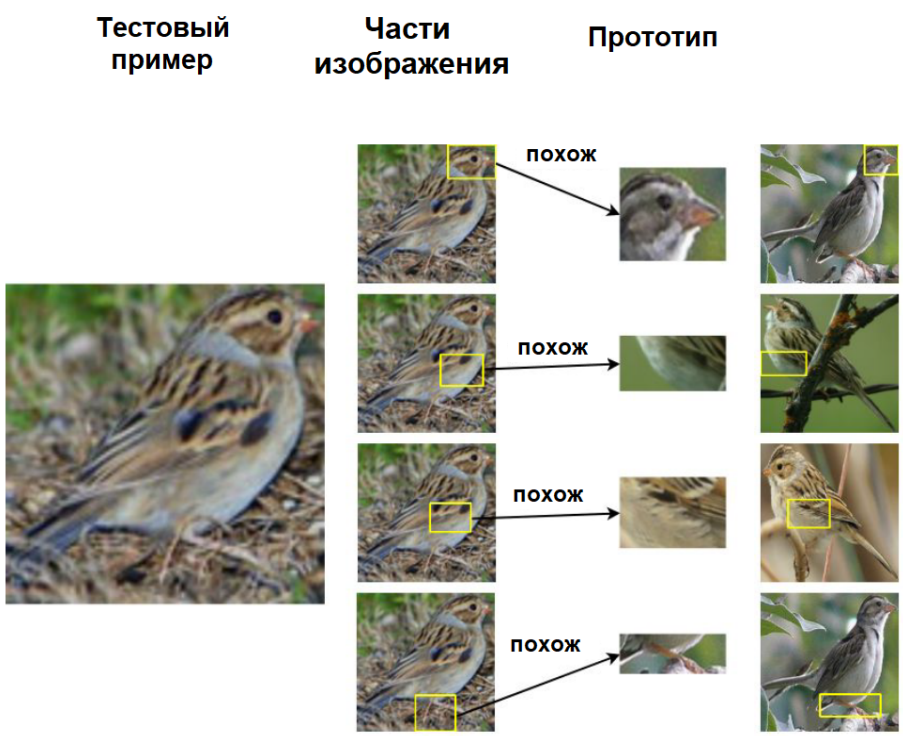
Объяснения примером (example-based explanation)

- Методы выбирают пример (не признаки) из датасета для объяснения поведения основной модели
- Имеют смысл, если можно представить пример данных в виде понятном человеку
- Используют **прототипы классов** в задачах классификации
- **k-ближайших соседей**: X классифицируется как y, так как A, B и C из y аналогичны X

Не лучший метод, так как объясняемый пример может быть далеко от объясняющего примера!

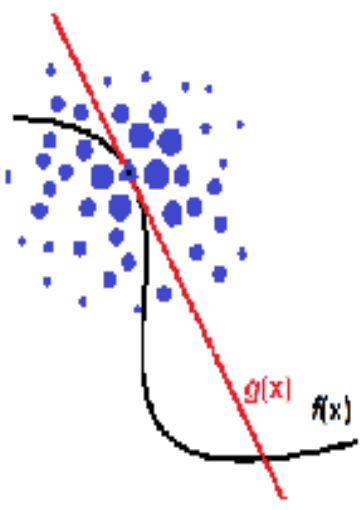
Прототипы классов (полные и фрагментарные)

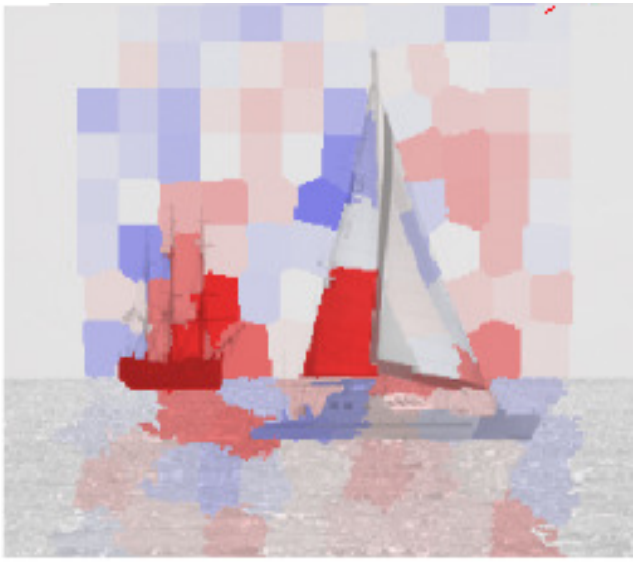




Метод возмущений (Perturbation)

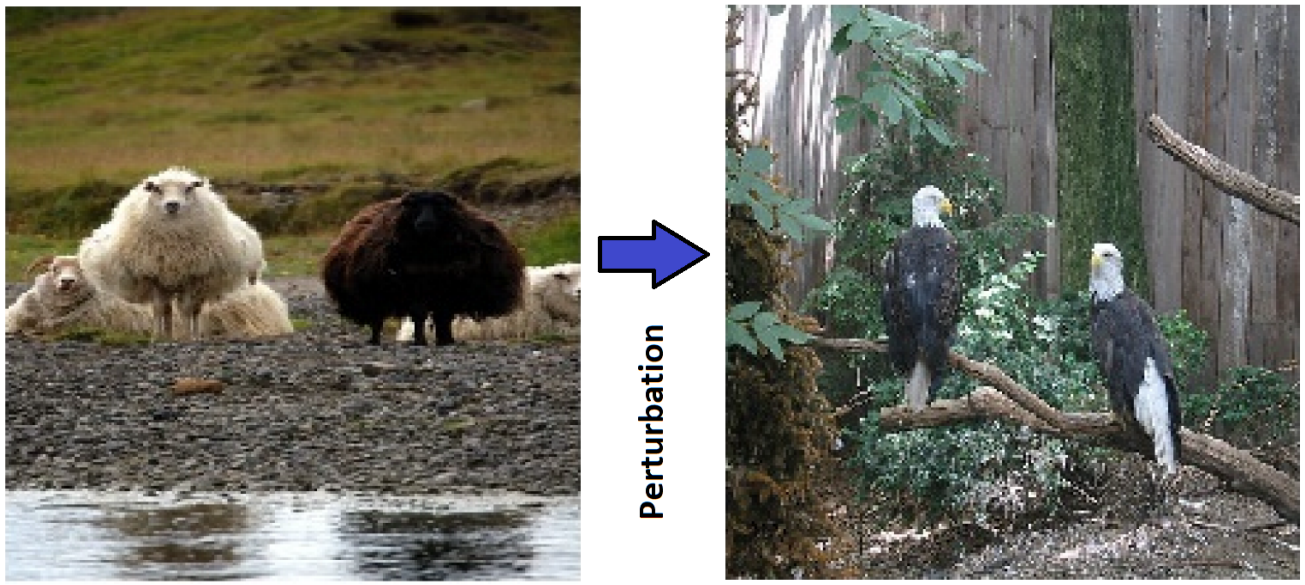
$\hat{\mathbf{x}} = \mathbf{x} + \delta$, δ - вектор возмущений (случайных или вычисляемых)





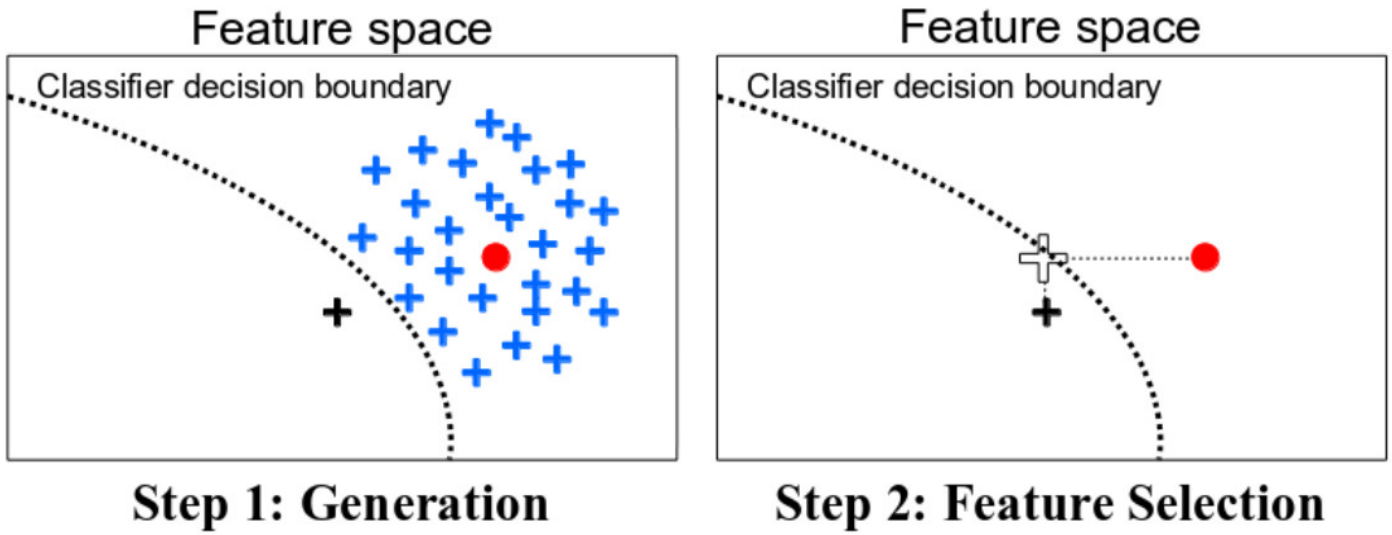
Метод возмущений

$\hat{\mathbf{x}} = \mathbf{x} + \delta$, δ - вектор возмущений (случайных или вычисляемых)



Counterfactual объяснения (гипотетические, противопоставления)

- Counterfactual - наименьшие изменения значений признаков, которые изменяют класс примера
- “Ваш запрос на кредит отклонен, так как ваш доход $30,000 и ваш баланс $200. Если бы ваш доход был $35,000 и ваш текущий баланс был $400, то ваш запрос был бы одобрен”



Counterfactuals (1)

- Обычно спрашивают, не почему был сделан определенный прогноз, а почему этот прогноз был сделан вместо другого прогноза.
- Для прогноза стоимости дома человека может интересовать, почему прогнозируемая цена была выше по сравнению с более низкой ценой, которую он ожидал.
- Когда заявка на кредит отклонена, меня не интересует, почему отказ. Меня интересуют факторы моей заявки, которые должны измениться, чтобы она была принята.
- Противоречивые объяснения легче понять, чем полные объяснения.

Counterfactuals (2)

- Врач задается вопросом: «Почему лечение не сработало на пациенте?»
- Полное объяснение, почему лечение не работает, включает: пациент болеет с 10 лет, 11 генов сверхэкспрессированы, что делает болезнь более тяжелой, организм пациента разрушается, лекарство неэффективно
- Сравнительное объяснение - отвечает на вопрос по сравнению с другим пациентом, для которого препарат работал, может быть проще: у пациента есть комбинация генов, которые делают лекарство неэффективным, по сравнению с другим пациентом
- Лучшее объяснение - это то, что подчеркивает наибольшую разницу между объектом интереса и “эталонным” объектом

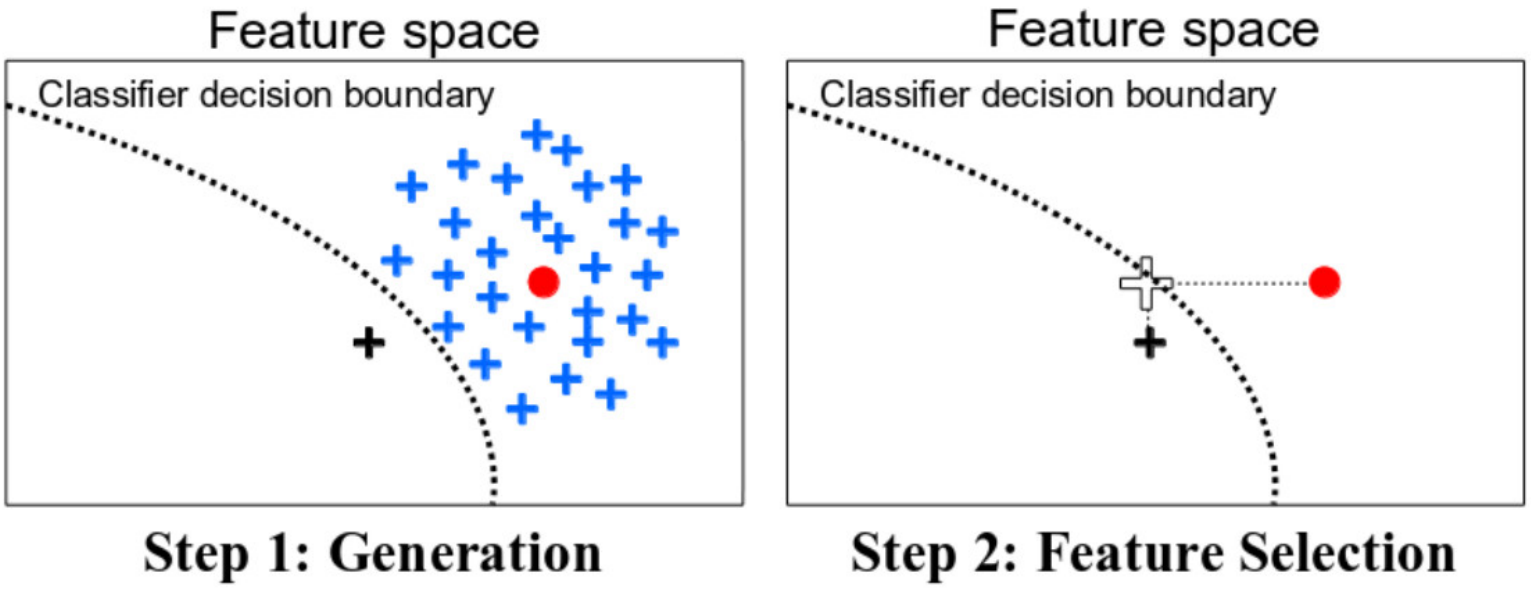
Counterfactuals (3)

- Counterfactuals - наименьшие изменения значений признаков, которые изменяют класс примера
- Counterfactual $\mathbf{z}$ для примера $\mathbf{x}$ определяется решением задачи оптимизации:

$$\min_{\mathbf{z} \in \mathbb{R}^m} L(f(\mathbf{z}), f(\mathbf{x})) + C\theta(\mathbf{z}, \mathbf{x})$$

  - $L(\cdot, \cdot)$ - функция потерь, устанавливающая связь между выходами основной модели;
  - $\theta(\cdot, \cdot)$ - штрафное слагаемое "против"больших отклонений $\mathbf{z}$ от $\mathbf{x}$, например, расстояние между $\mathbf{z}$ и $\mathbf{x}$;
  - $C > 0$ - параметр

Counterfactuals (4)



Глобальная интерпретация - feature importance

- Значимость признаков - какие признаки оказывают наибольшее влияние на прогнозируемые значения?
- Алгоритм: значимость перестановок (permutation importance)
  1. Получить обученную модель и записать признаки в виде таблицы (столбец - признак)
  2. Перемешать значения в одном столбце, сделать прогнозы, используя полученный набор данных. Снижение точности - значимость признака, который перемешали.
  3. Вернуться к исходной таблице (отмена перемешивания из шага 2). Повторить шаг 2 со следующим столбцом в таблице, пока не будут найдены значимости каждого столбца.

Глобальная интерпретация - Partial Dependence Plot

- **Значимость признаков** показывает, **какие** признаки больше всего влияют на прогнозы, **график частичной зависимости** показывает, **как** признак влияет на прогнозы
- График частичной зависимости показывает, какая зависимость между признаком и выходом: линейнеая, монотонная или более сложная

График частичной зависимости (пример 1)

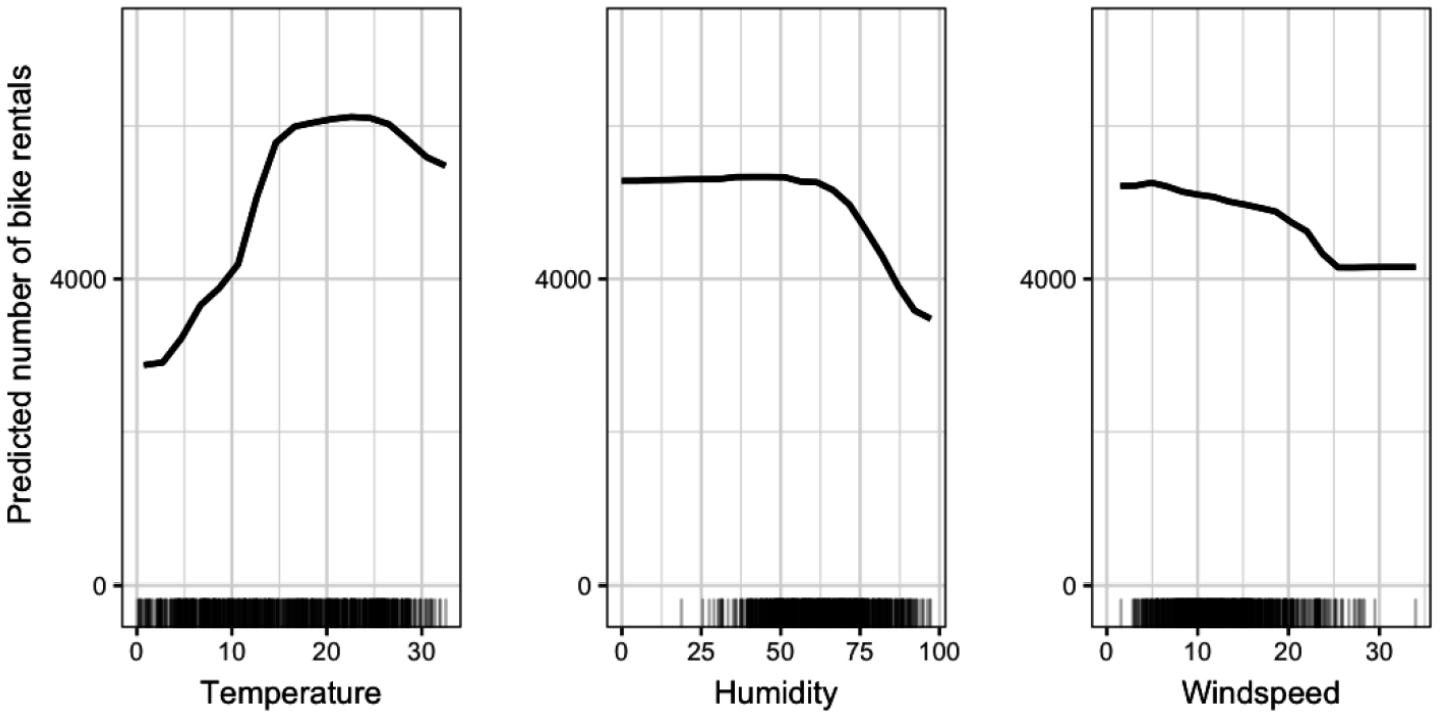


График частичной зависимости (пример 2 - взаимодействие признаков)

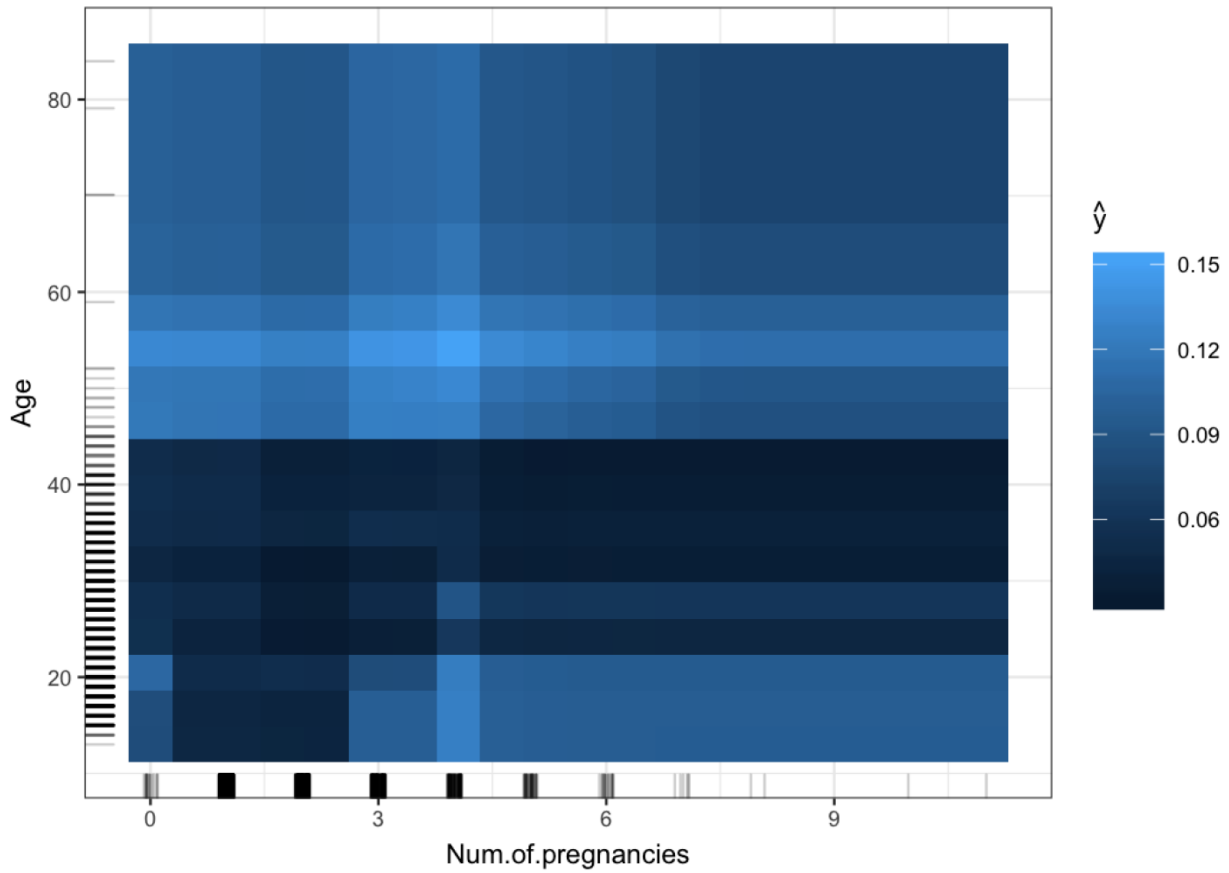


График частичной зависимости

- Частичная зависимость:

$$f_{x_S}(x_S) = \mathbb{E}_{x_C}\left[f(x_S, x_C)\right] = \int f(x_S, x_C) d\mathbb{P}(x_C)$$

- x_S - множество признаков, для которых график частичной зависимости определяется
- x_C - все другие признаки, используемые в модели f; $x = x_S || x_C$ (конкатенация)
- Частичная зависимость работает путем маргинализации выходных данных модели f по распределению признаков x_C, так что оставшаяся функция показывает связь между x_S и прогнозом

График частичной зависимости

- Частичная зависимость по данным из датасета:

$$f_{x_S}(x_S) = \frac{1}{n}\sum_{i=1}^{n} f(x_S, x_{C_i})$$

- x_{C_i} - фактические значения признаков из датасета, в которых мы не заинтересованы
- Используемое предположение: признаки x_S не коррелируют с признаками x_C

Чтобы закончить...

"Look deep into nature, and then you will understand everything better."
Albert Einstein

Вопросы

Вопросы —– ?