Анализ данных и Машинное обучение в гидрологии

...

Неделя 5

План

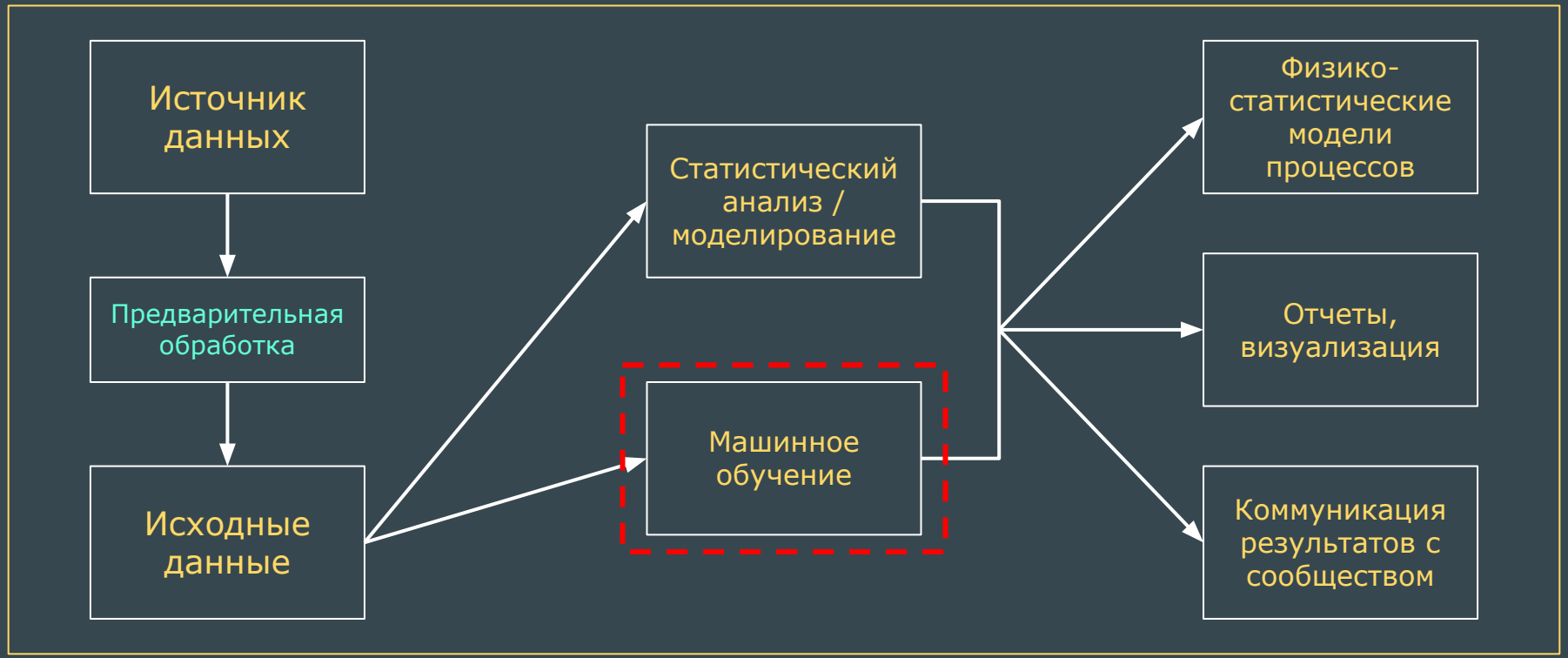
Лекция

- что такое машинное обучение?
- сможет ли машинное обучение решить все проблемы?
- могу ли я доверять решениям машины?
- гугл строит новый скайнет. Когда мы все умрем?

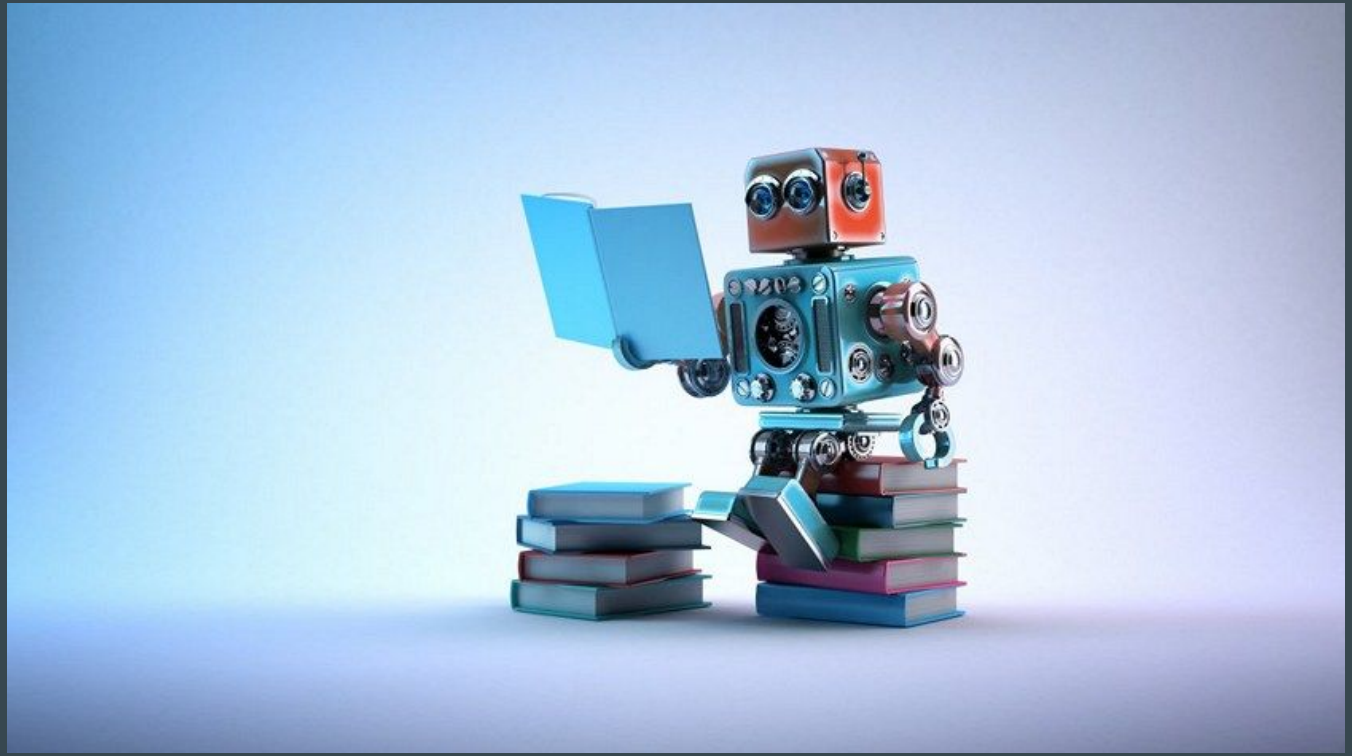
Вопросы

- какие планы на год?
- какая модель машинного обучения самая лучшая?
- я скоро заканчиваю универ, стоит мне идти в науку?
- куда переедет ИВП РАН?
- когда следующий **Plovcast**?

Research workflow

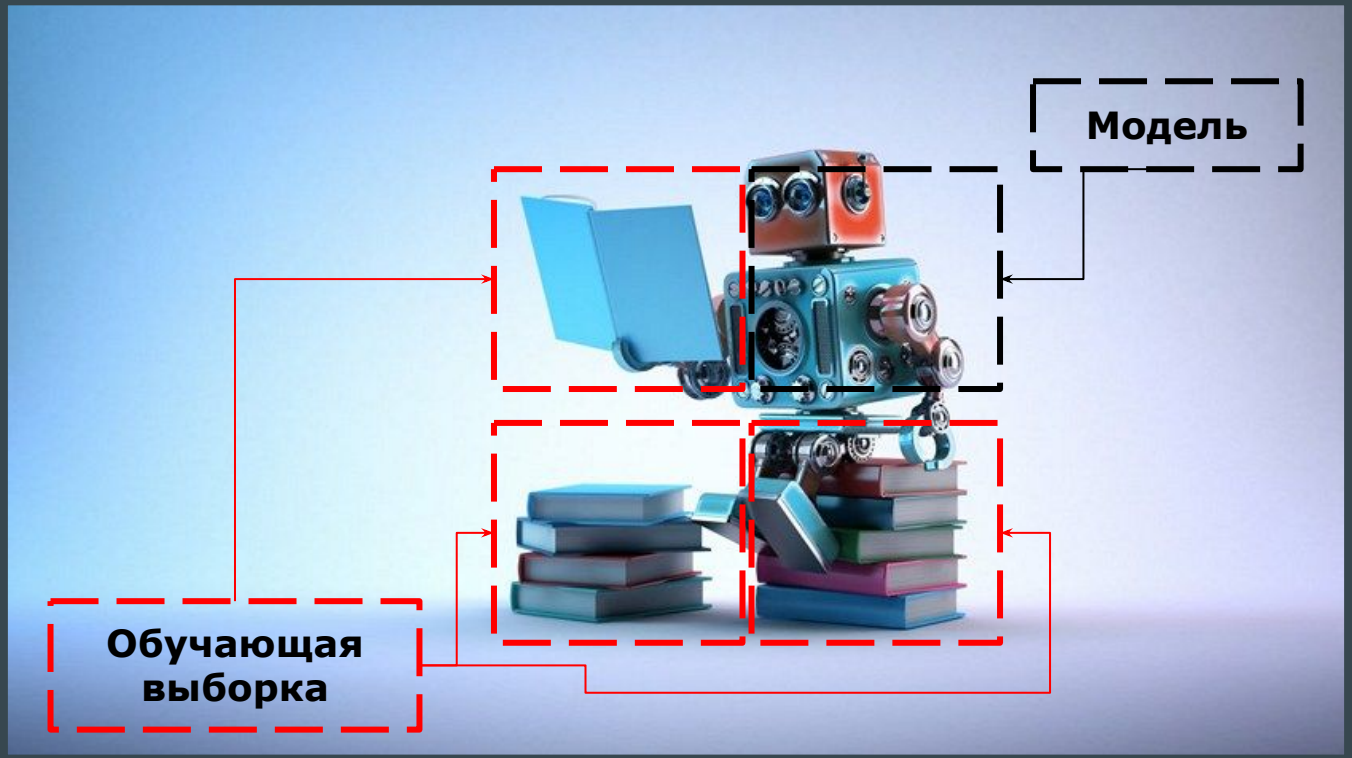


Машинное обучение



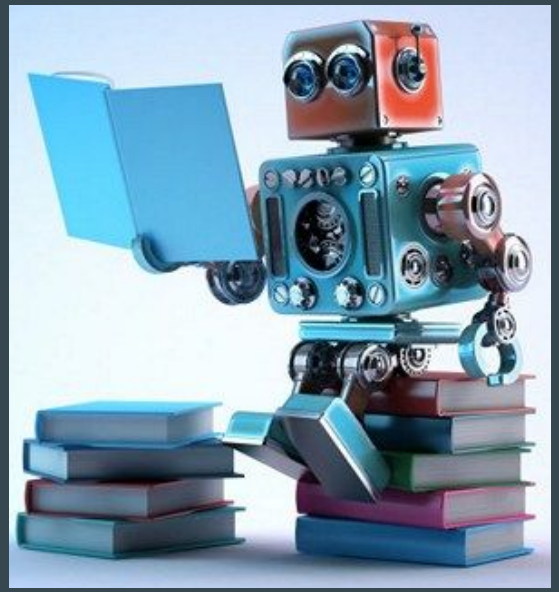
http://gagadget.com/media/post_big/books_ai.jpg

Машинное обучение



http://gagadget.com/media/post_big/books_ai.jpg

Машинное обучение

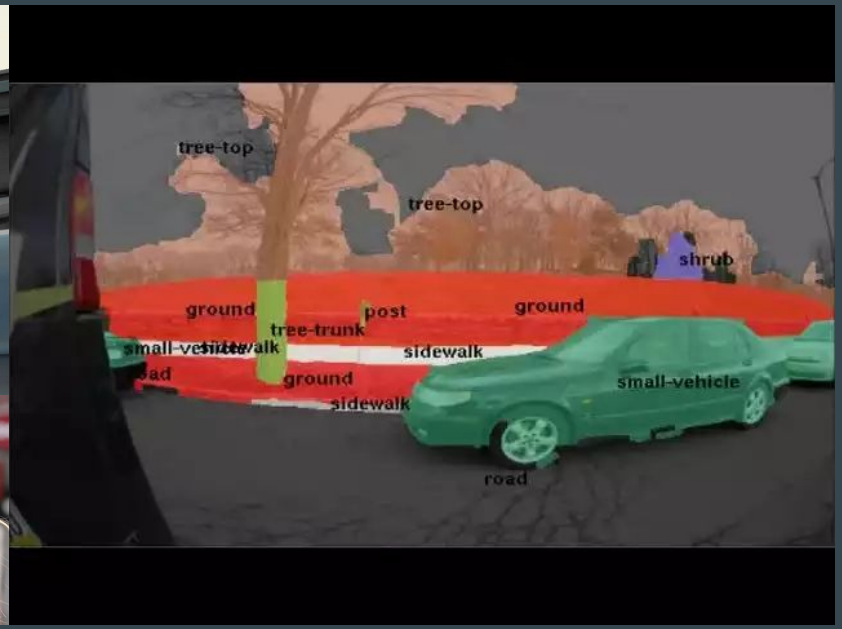


ИВП РАН

Google, Yandex

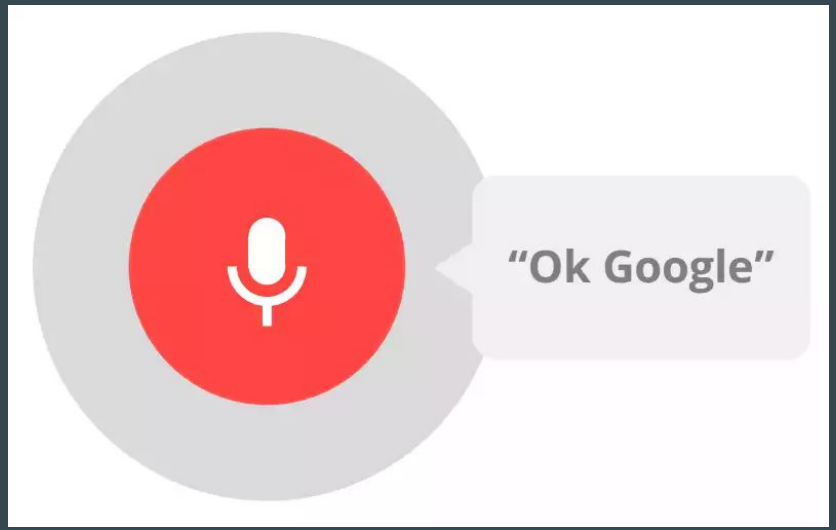
Что умеют современные модели

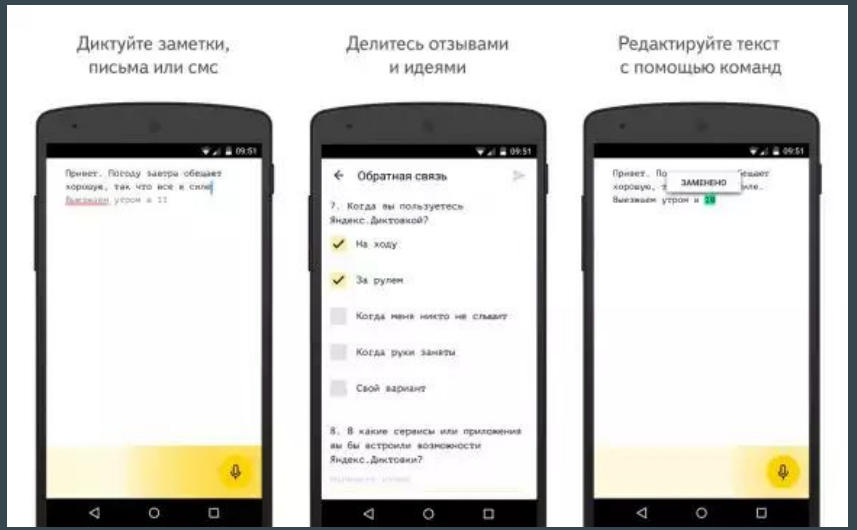
Распознавание образов



Что умеют современные модели

Распознавание речи

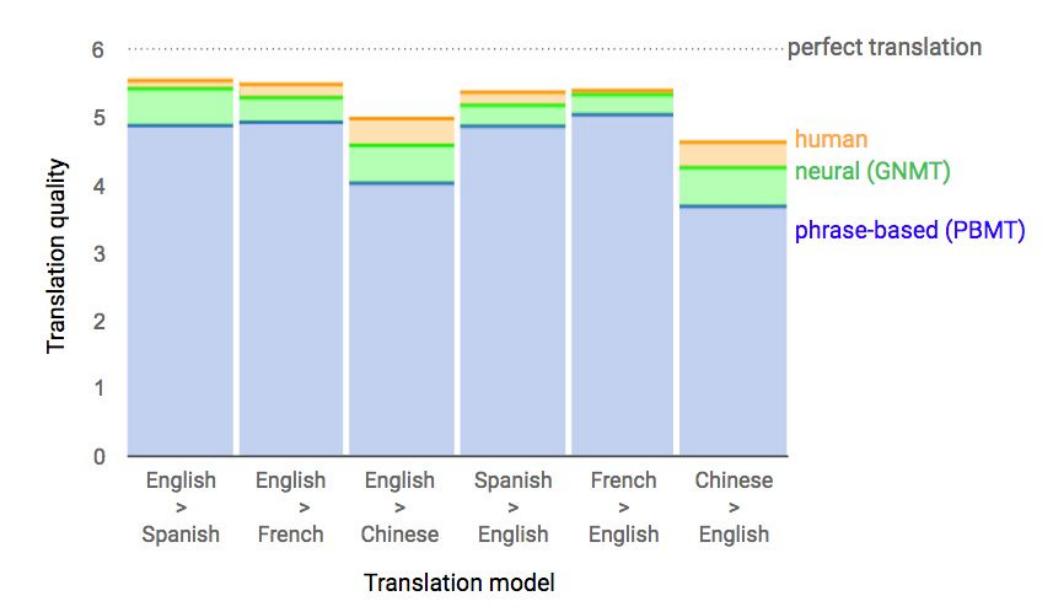




Яндекс.Диктовка

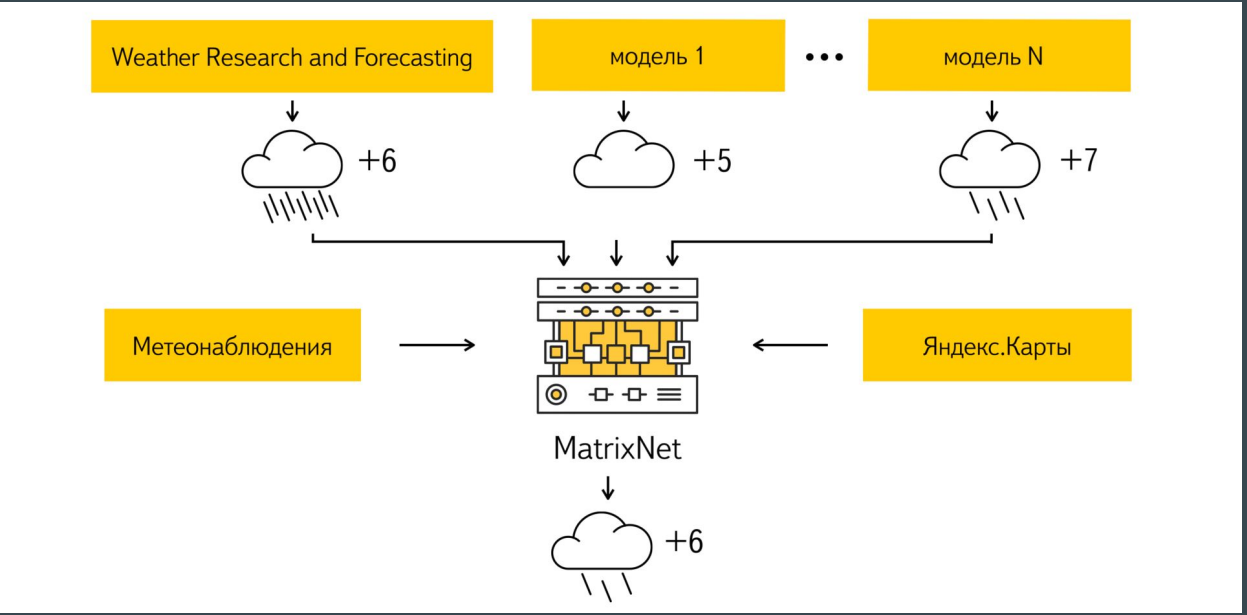
Что умеют современные модели

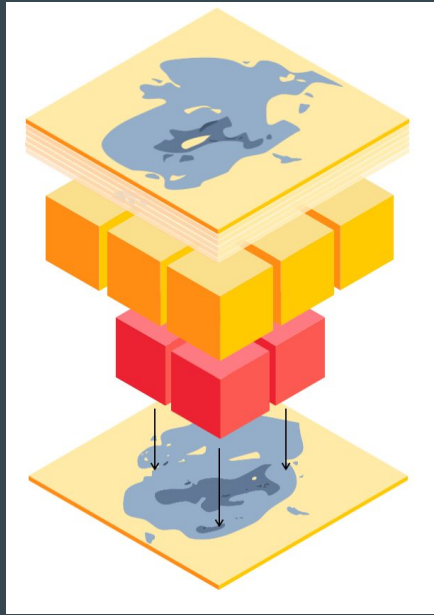
Перевод



Что умеют современные модели

Улучшение прогноза погоды





Но как?

КАК ОНИ ЭТО ДЕЛАЮТ?

Ну как-то так...

$$
\begin{aligned}
&\propto \frac{\prod_{i\neq k}\Gamma(n_{m,(\cdot)}^{i,-(m,n)}+\alpha_i)}{\Gamma\big((\sum_{i=1}^{K} n_{m,(\cdot)}^{i,-(m,n)}+\alpha_i)+1\big)} \prod_{i\neq k} \frac{\Gamma(n_{(\cdot),v}^{i,-(m,n)}+\beta_v)}{\Gamma\big(\sum_{r=1}^{V} n_{(\cdot),r}^{i,-(m,n)}+\beta_r\big)} \\
&\times \Gamma(n_{m,(\cdot)}^{k,-(m,n)}+\alpha_k+1) \frac{\Gamma(n_{(\cdot),v}^{k,-(m,n)}+\beta_v+1)}{\Gamma\big((\sum_{r=1}^{V} n_{(\cdot),r}^{k,-(m,n)}+\beta_r)+1\big)} \\
&\propto \frac{\Gamma(n_{m,(\cdot)}^{k,-(m,n)}+\alpha_k+1)}{\Gamma\big((\sum_{i=1}^{K} n_{m,(\cdot)}^{i,-(m,n)}+\alpha_i)+1\big)} \frac{\Gamma(n_{(\cdot),v}^{k,-(m,n)}+\beta_v+1)}{\Gamma\big((\sum_{r=1}^{V} n_{(\cdot),r}^{k,-(m,n)}+\beta_r)+1\big)} \\
&= \frac{\Gamma(n_{m,(\cdot)}^{k,-(m,n)}+\alpha_k)\big(n_{m,(\cdot)}^{k,-(m,n)}+\alpha_k\big)}{\Gamma\big(\sum_{i=1}^{K} n_{m,(\cdot)}^{i,-(m,n)}+\alpha_i\big)\big(\sum_{i=1}^{K} n_{m,(\cdot)}^{i,-(m,n)}+\alpha_i\big)} \frac{\Gamma\big(n_{(\cdot),v}^{k,-(m,n)}+\beta_v\big)\big(n_{(\cdot),v}^{k,-(m,n)}+\beta_v\big)}{\Gamma\big(\sum_{r=1}^{V} n_{(\cdot),r}^{k,-(m,n)}+\beta_r\big)\big(\sum_{r=1}^{V} n_{(\cdot),r}^{k,-(m,n)}+\beta_r\big)} \\
&\propto \frac{\big(n_{m,(\cdot)}^{k,-(m,n)}+\alpha_k\big)}{\big(\sum_{i=1}^{K} n_{m,(\cdot)}^{i,-(m,n)}+\alpha_i\big)} \frac{\big(n_{(\cdot),v}^{k,-(m,n)}+\beta_v\big)}{\big(\sum_{r=1}^{V} n_{(\cdot),r}^{k,-(m,n)}+\beta_r\big)} \\
&\propto \big(n_{m,(\cdot)}^{k,-(m,n)}+\alpha_k\big) \frac{\big(n_{(\cdot),v}^{k,-(m,n)}+\beta_v\big)}{\big(\sum_{r=1}^{V} n_{(\cdot),r}^{k,-(m,n)}+\beta_r\big)}.
\end{aligned}
$$

Взгляд со стороны

Отношение к машинному обучению

Нужно принять решение

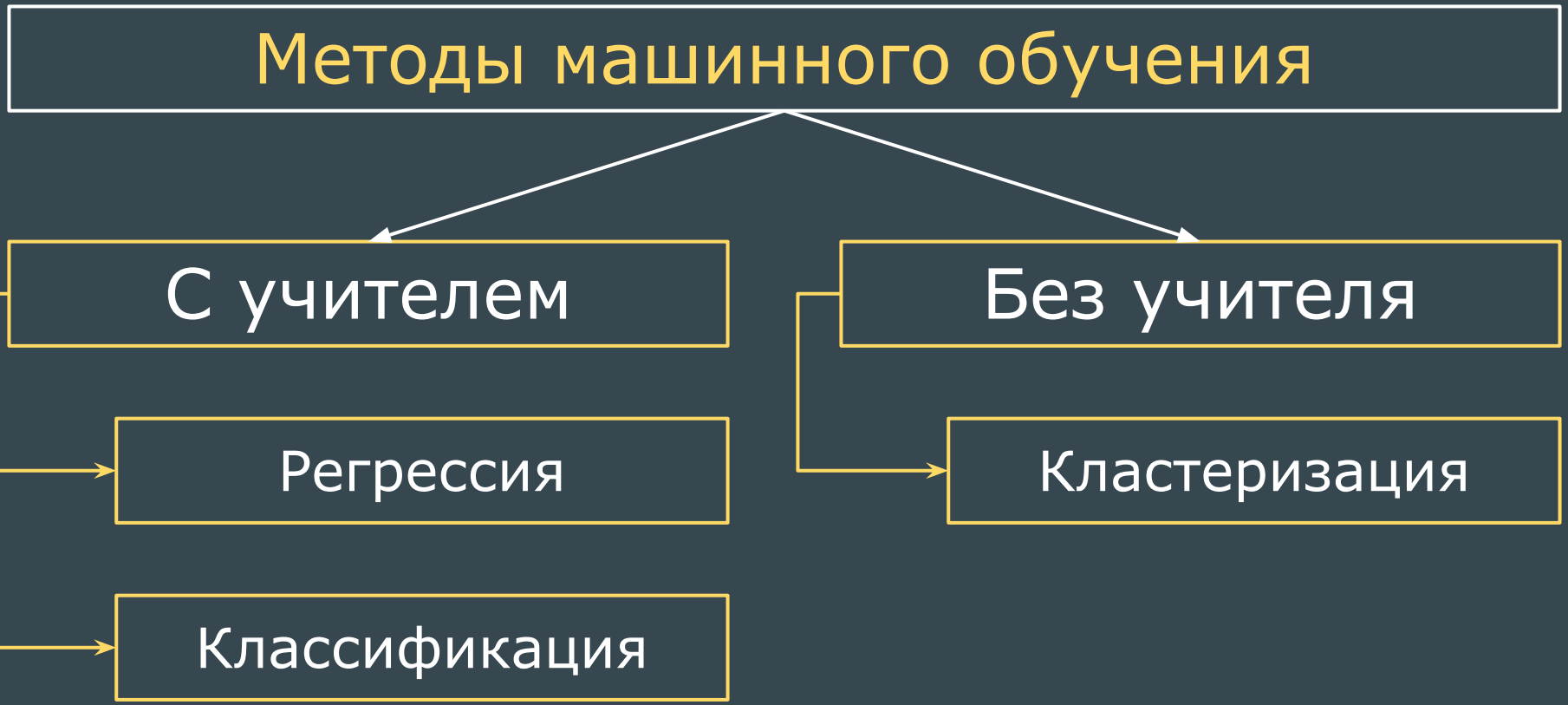
Разобраться

- Время: > месяца
- Усилия: большие

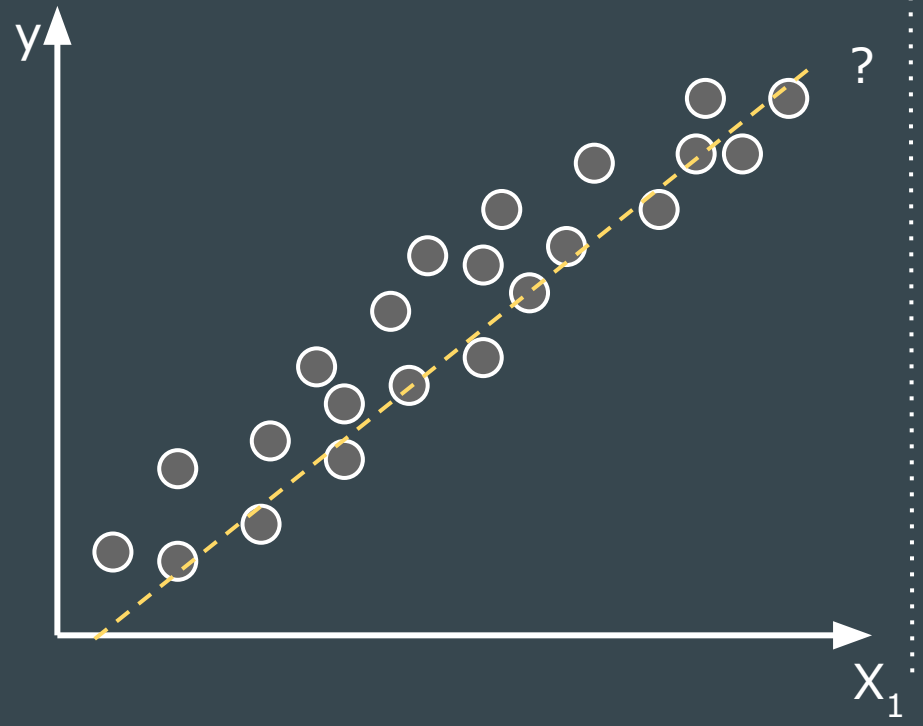
Сказать, что это говно

- Время: сразу
- Усилия: никаких

МАТАН — ДОБРО


Методы машинного обучения
С учителем
Без учителя
Регрессия
Кластеризация
Классификация

Регрессия



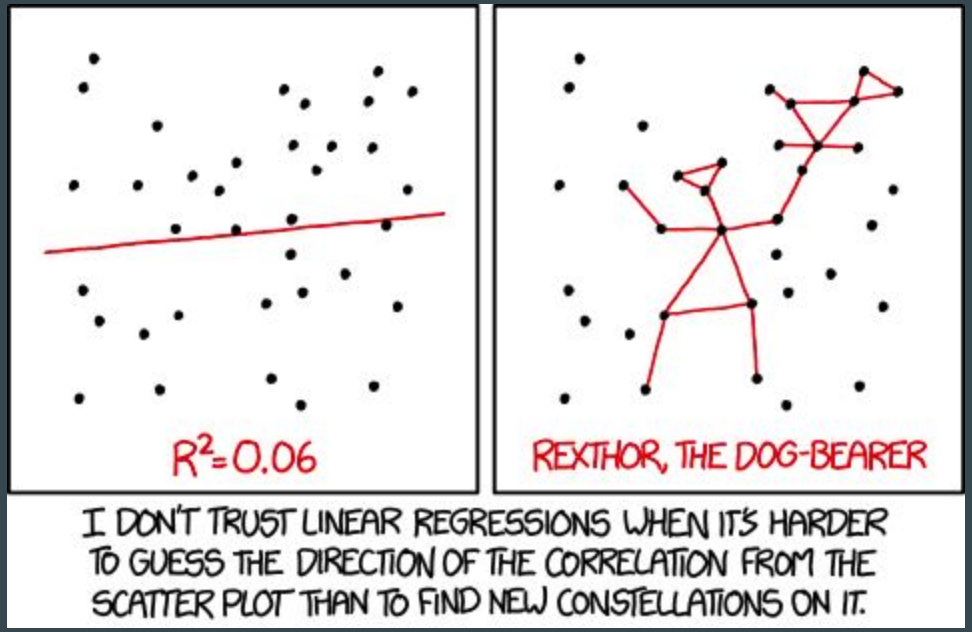
$$h_\theta(x) = \theta_0 + \theta_1 x$$

$$\theta_0, \theta_1$$

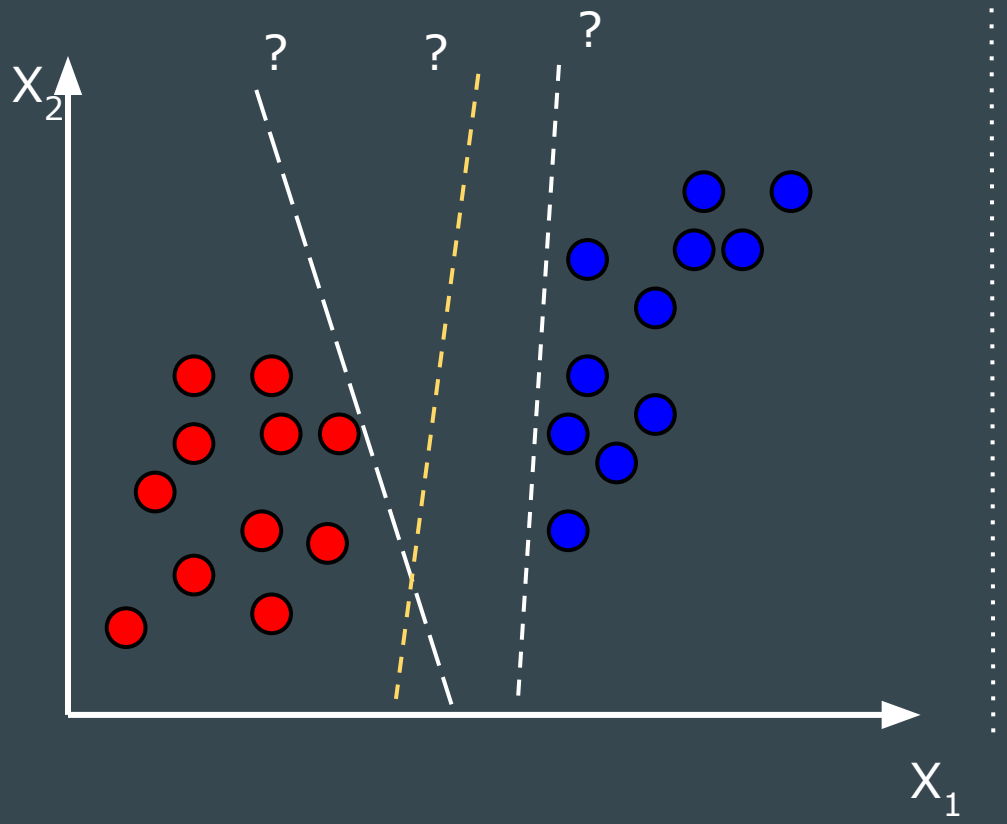
$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

$$\underset{\theta_0, \theta_1}{\text{minimize}}\ J(\theta_0, \theta_1)$$

Регрессия


R^2=0.06
REXTHOR, THE DOG-BEARER
I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

Классификация



$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

$$\theta$$

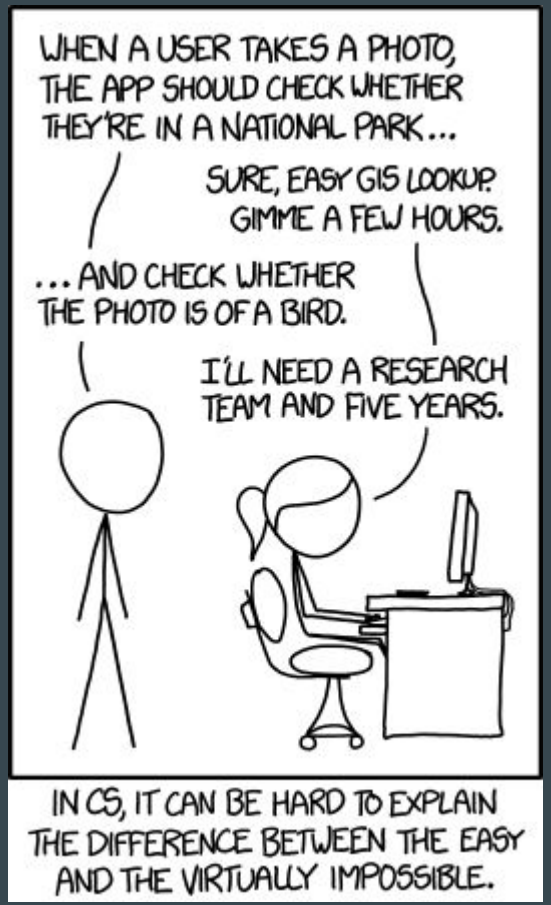
$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \text{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

$$\text{Cost}(h_\theta(x), y) =$$

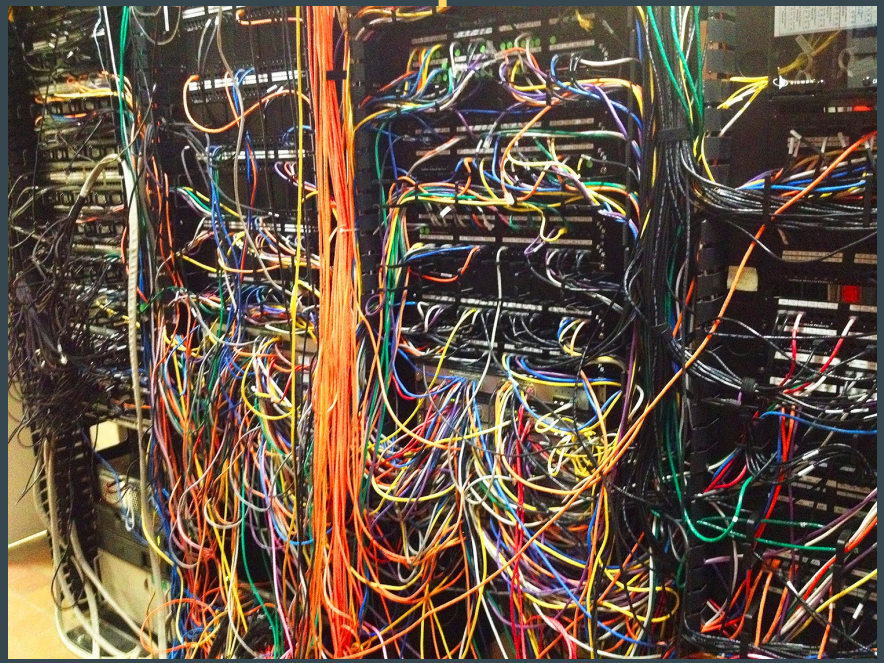
$$= \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

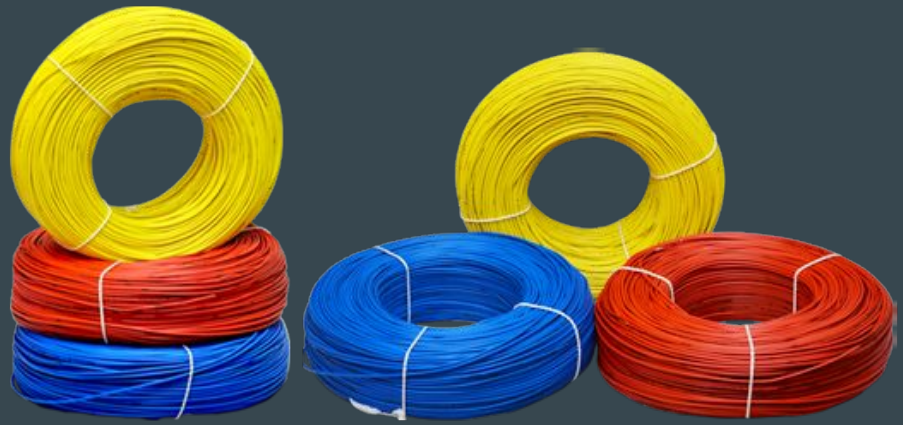
$$\min_\theta J(\theta)$$

Классификация

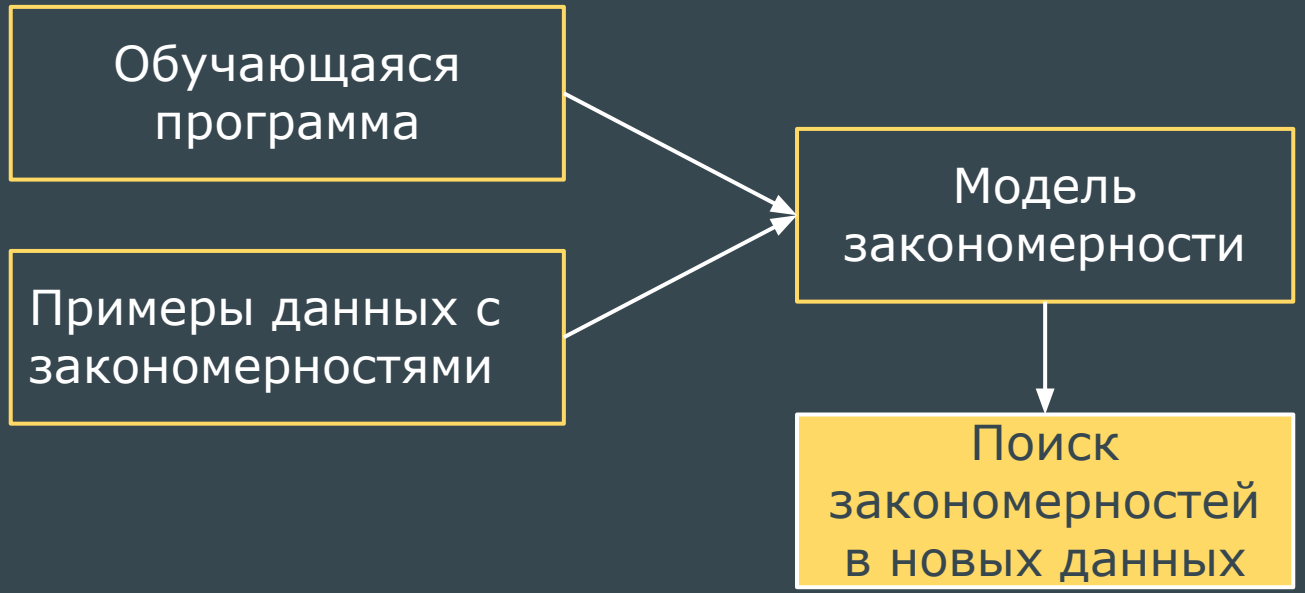


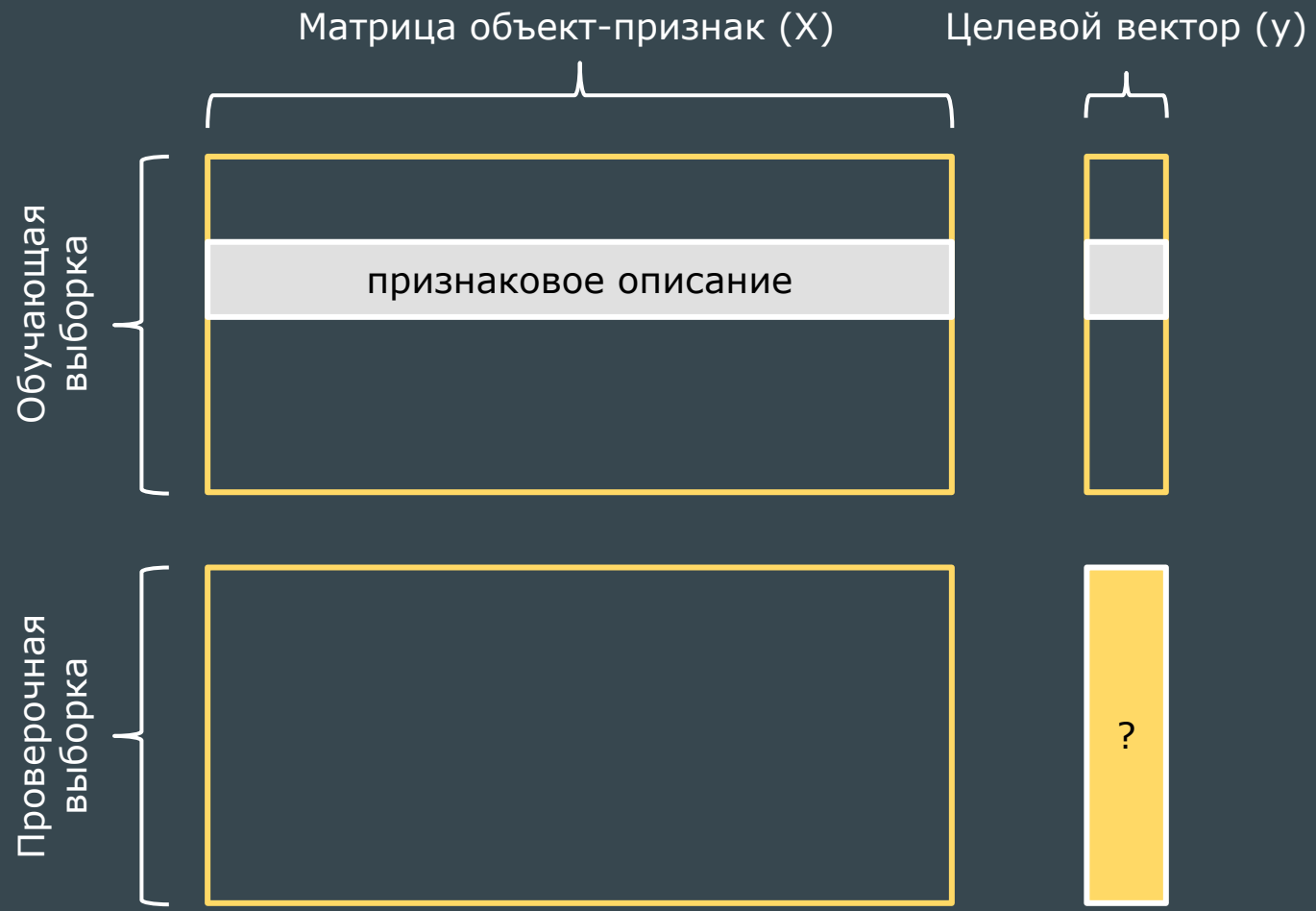
Кластеризация

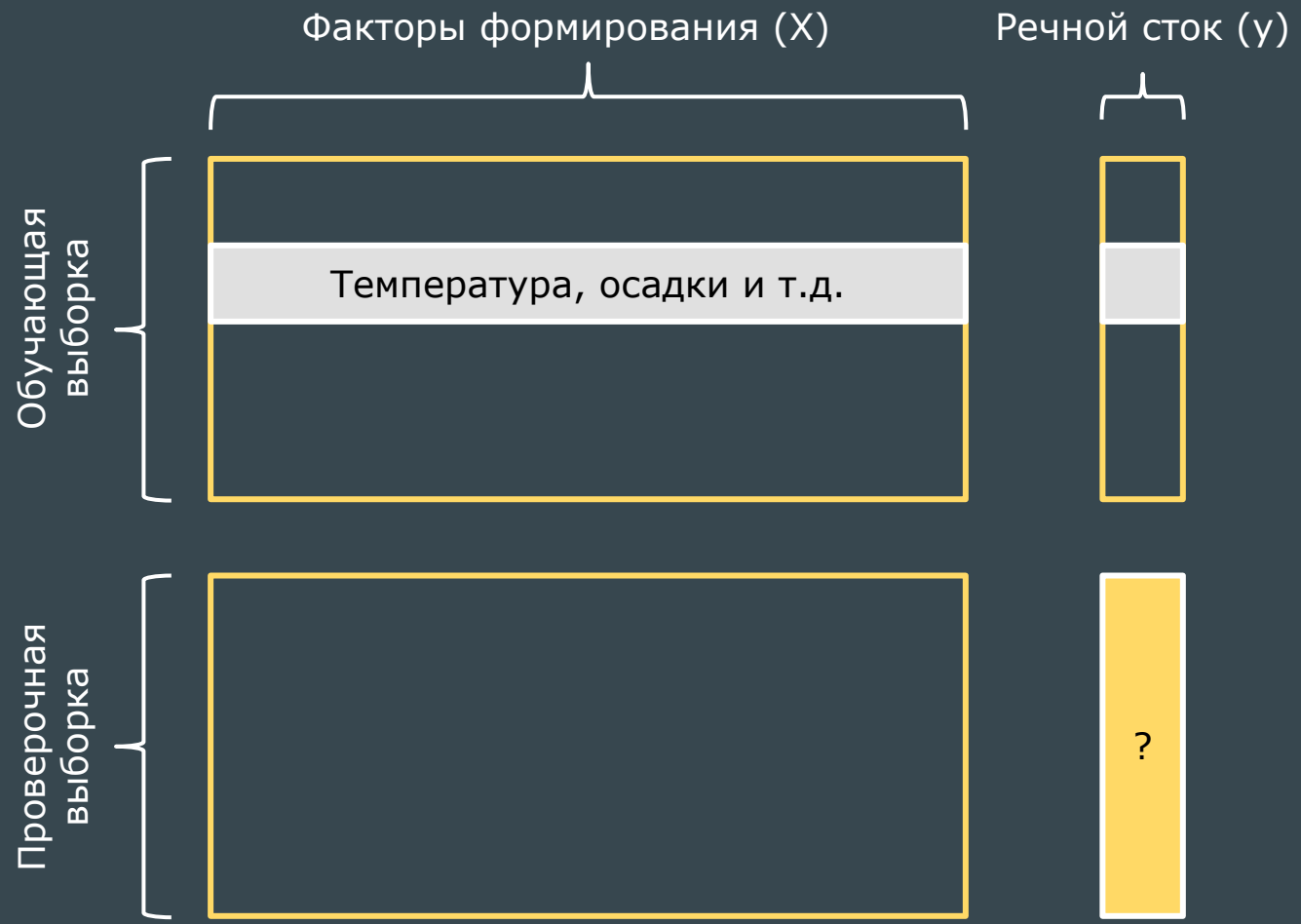




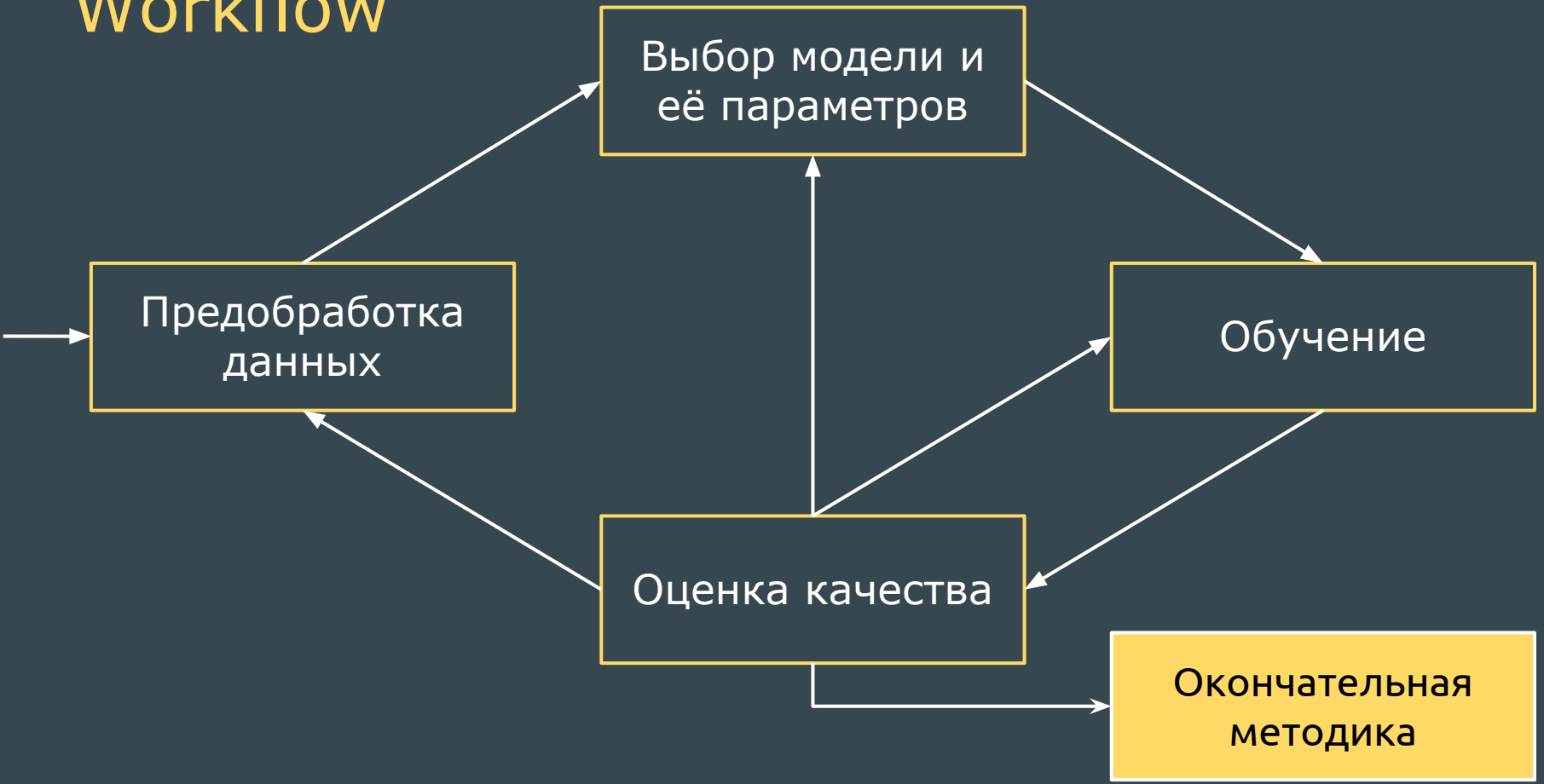
Идея обучения с учителем




Матрица объект-признак (X)
Целевой вектор (y)
Обучающая выборка
признаковое описание
Проверочная выборка
?


Факторы формирования (X)
Речной сток (y)
Обучающая выборка
Температура, осадки и т.д.
Проверочная выборка
?

Workflow


Выбор модели и её параметров
Предобработка данных
Обучение
Оценка качества
Окончательная методика

Workflow

Forming →

Norming →

Storming →

Performing

Forming (получение данных)

- txt
- csv
- netcdf
- sql
- xml
- web api

1. import …
2. path =
3. connection =
4. data = parse(path)

profit!

Norming (предварительная обработка)

- сортировка
- группировка
- заполнение пропусков
- удаление выбросов
- создание новых переменных

1. import numpy as np
2. import pandas as pd
3. from sklearn import Preprocessing
4. library.method()

profit!

Storming (моделирование, анализ)

- классификация
- кластеризация
- регрессия
- распознавание образов
- моделирование
- прогнозирование

1. from sklearn import SVR
2. model = SVR()
3. model.fit(X, y)
4. metrics(model)
5. model.predict(y)

profit!

Performing (представление результатов)

- научная графика
- воспроизводимые результаты
- переиспользование кода
- создание веб-приложений

- ❑ Matplotlib, Seaborn
- ❑ Ipython notebook, Docker, Git(hub)
- ❑ OOP, Gist
- ❑ Flask

Что дальше?

Стадии развития методов машинного обучения (Vapnik, 1995):

- первые алгоритмы машинного обучения
- основы теории
- нейронные сети
- альтернативы нейронным сетям
- глубокие нейронные сети (Айзель, 201X)

План

23.01	Нейронные сети
06.02	Альтернативы нейронным сетям
20.02	Глубокие нейронные сети
06.03 (+ 20.03)	Хакатон “Машинное обучение для расчетов речного стока”

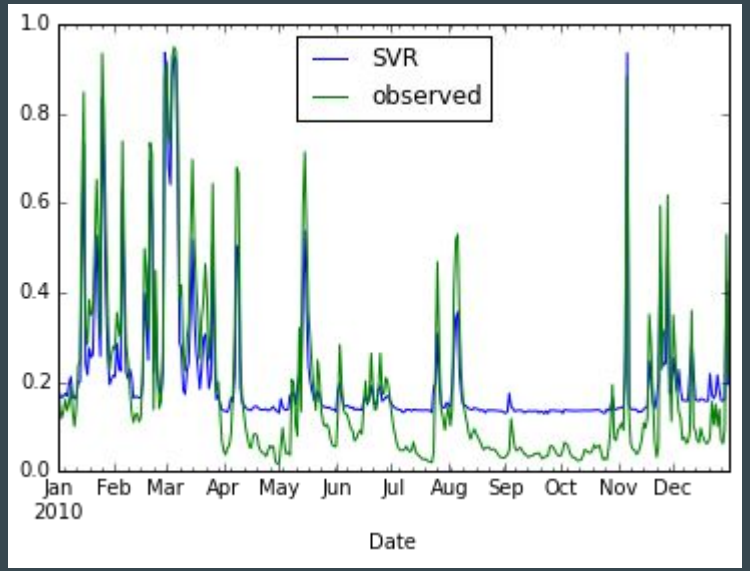
Чем занять себя две недели?

Микро-курс:

Getting started with machine learning in hydrology

Виноградовские чтения, 2015

Расчет паводочного стока малого водосбора с использованием машин опорных векторов



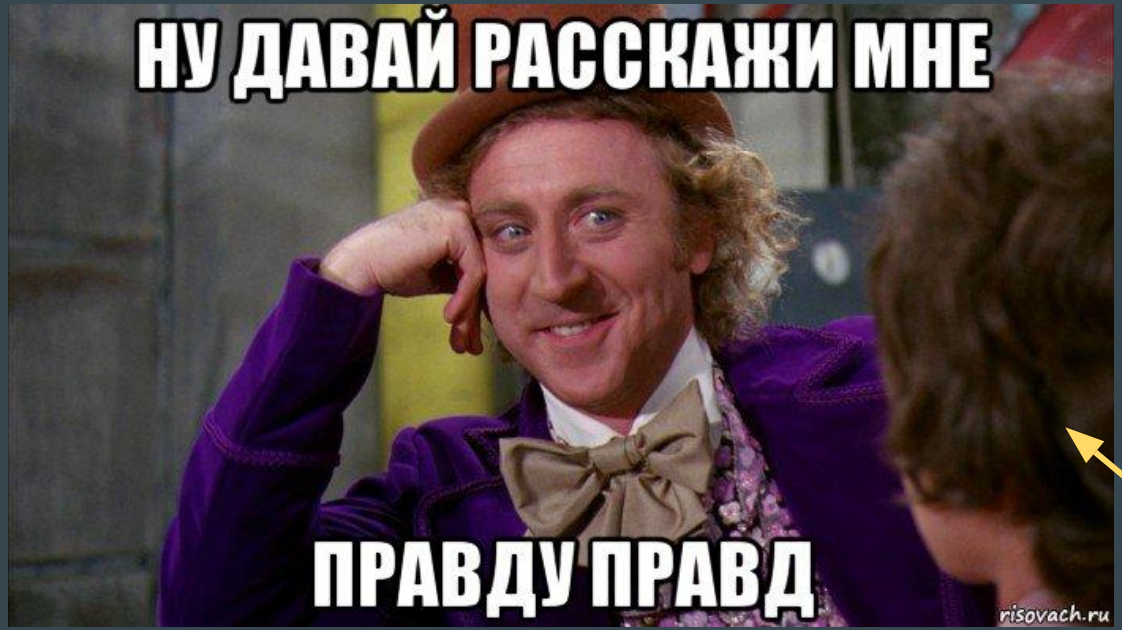
Важно

Вы можете помочь существенно улучшить этот курс!

- ayzelgv@gmail.com, hydrogo@yandex.ru
- vk.com/ayzelgv, facebook.com/ayzelgv
- ИВП РАН, кабинет 617

Q&A: questions and answers

AMA: ask me anything



Я