


1. Введение в машинное обучение



Зеленков Ю. А. (с) 2024

Содержание

- Краткое введение
- Виды данных и основные задачи.
- Инфраструктура машинного обучения.
- Статистические методы обучения на табличных данных.
- Предиктивные модели. Линейная регрессия.
- Предиктивные модели. Бинарная классификация.
- Экспланаторный анализ данных. Кластеризация

Оценим ваш бэкграунд



<https://forms.gle/uUzGYChijJMVdvXo8>

Содержание лекций

- 6.03: введение, линейные модели
- 13.03: деревья, нейронные сети, ансамбли, метрики
- 20.03: процесс решения задачи ML
- 27.03: глубокое обучение
- 29.03:
 - Обучение с подкреплением
 - AutoML
 - Интерпретация моделей
 - Каузальное моделирование

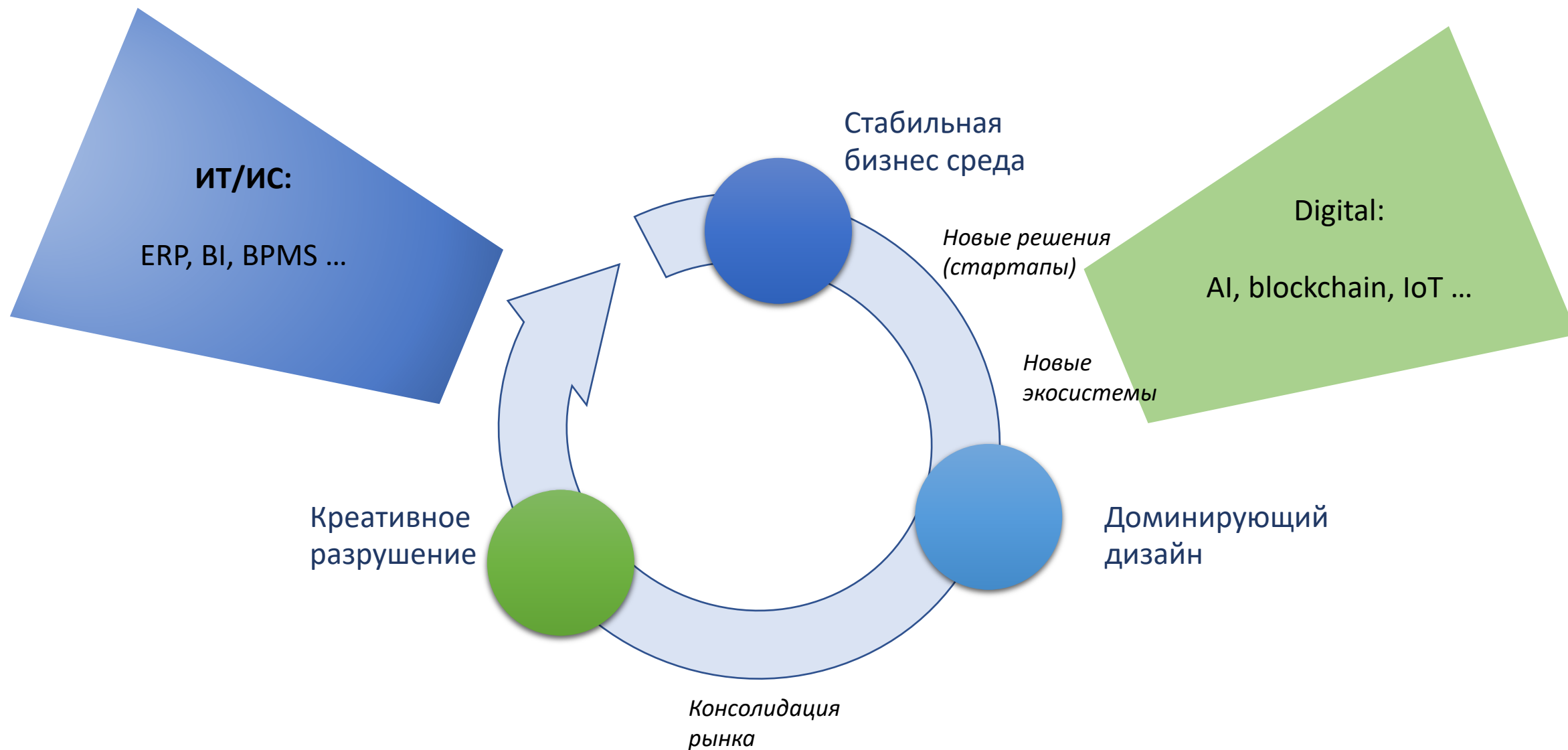
Рекомендуемая литература

- Флах, П. (2015) Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных. ДМК Пресс.
- Шолле, Ф. (2022). Глубокое обучение на Python. Питер.
- Мишра, П. (2022). Объяснимые модели искусственного интеллекта на Python. ДМК Пресс.
- Рассел, С., & Норвиг, П. (2016). Искусственный интеллект: Современный подход. Вильямс.

Материалы курса

- https://github.com/yzelenkov/ML_for_MBI/

Цикл технологического развития

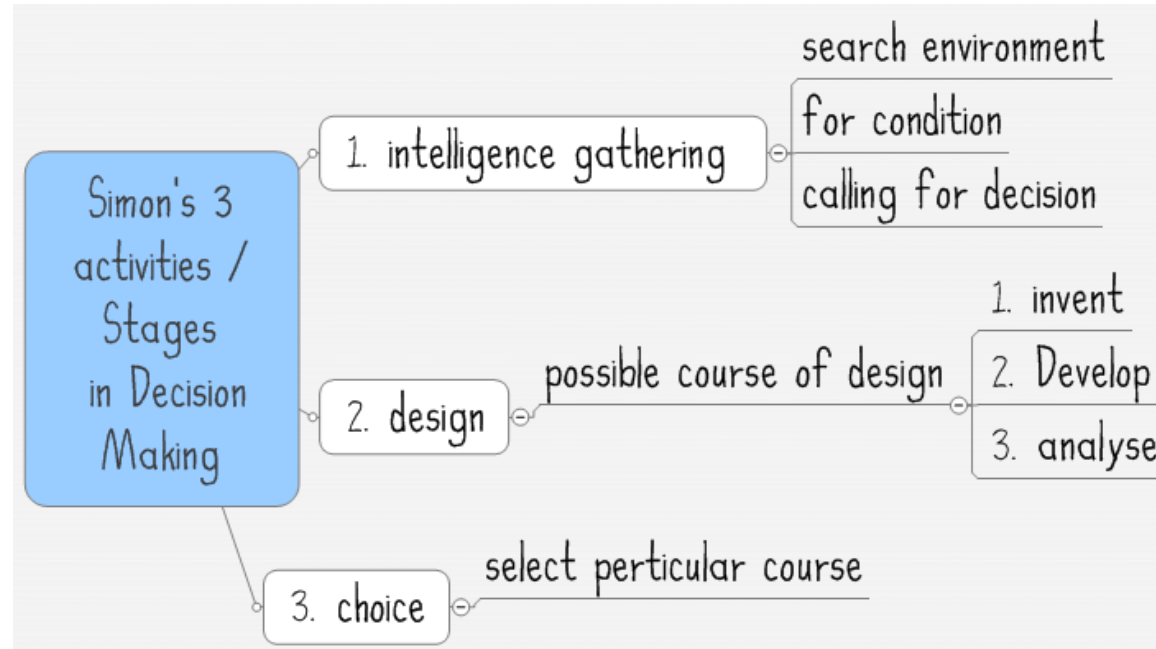


Процесс принятия решений



Герберт Саймон
(1916 - 2001)

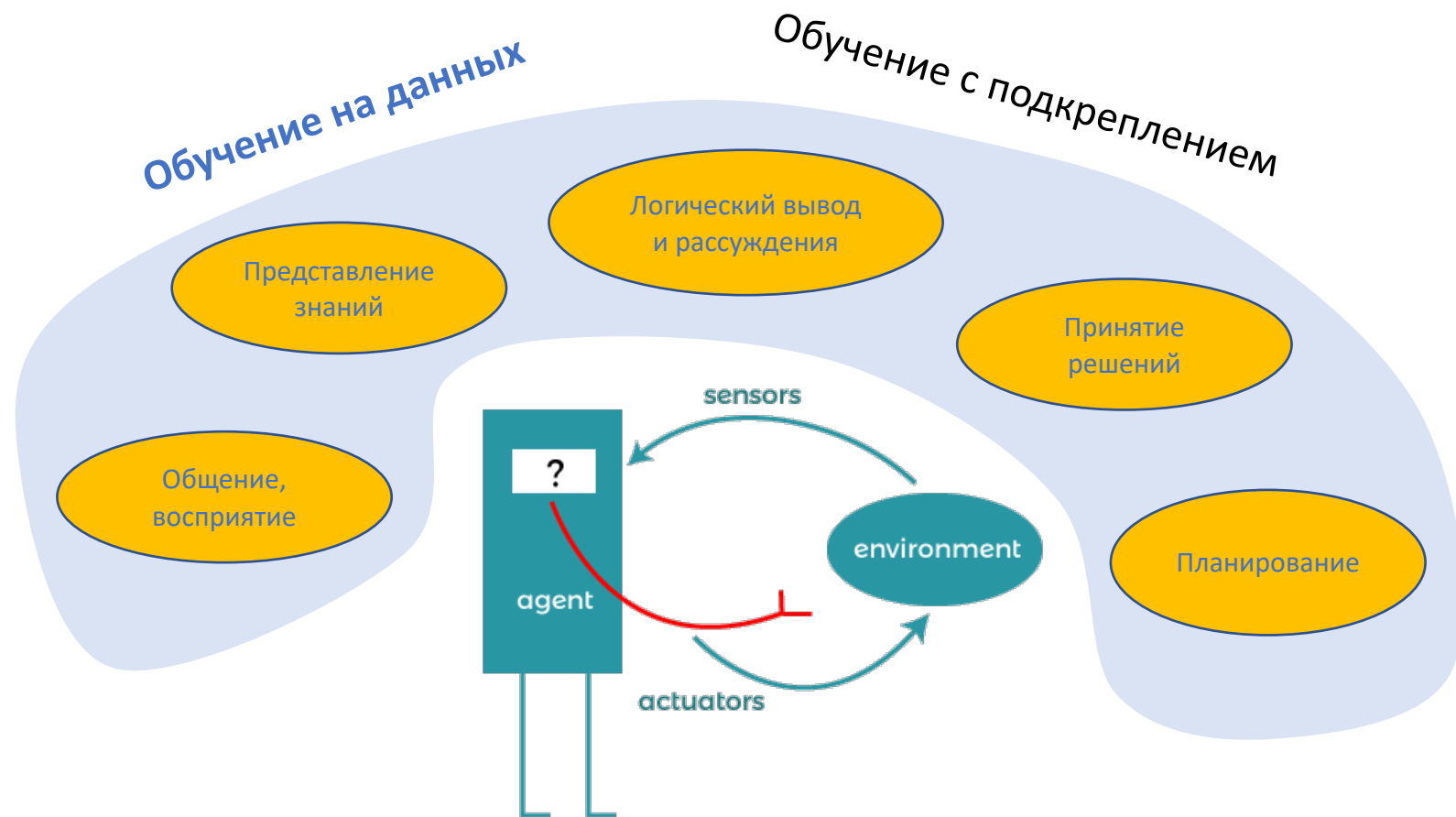
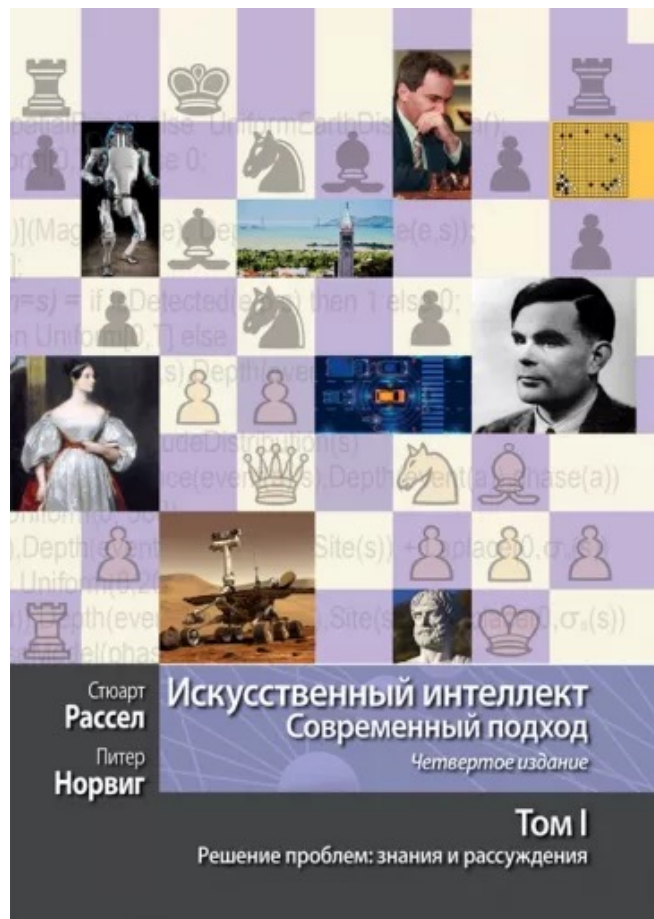
Премия Тьюринга – 1975
Нобелевская премия - 1978



Garbage Can Model



Искусственный интеллект* как наука



* термин AI введен Дж. Маккарти в 1956 г.

Виды данных

Структурированные

Неструктурированные

Табличные
(кросс-секционные)

Временные ряды




Панельные

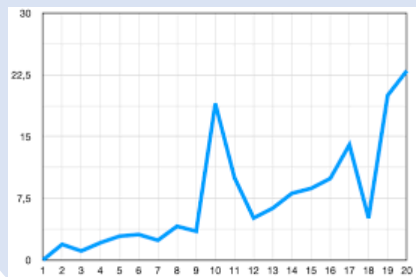
время

время

атрибуты

объекты

атрибуты			
	M	182	89
	F	165	58
	M	176	



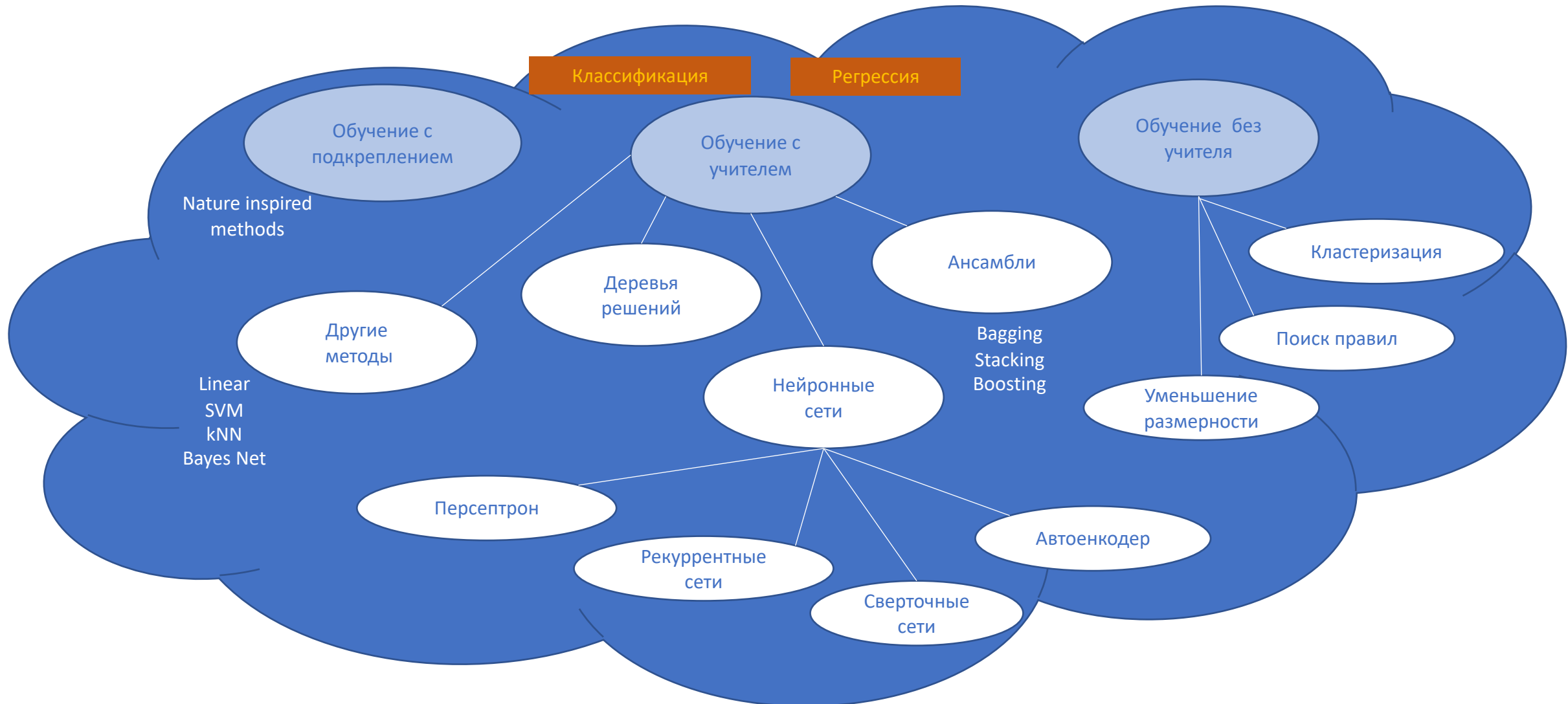
Text

Image

Video

Sound

Machine Learning





Инфраструктура
ML



TF 2.0 2019

2015

tensorflow

библиотека машинного обучения

IPython

интерактивная оболочка

sci-kit learn

python библиотека машинного обучения

NumPy / SciPy

python библиотеки для научных расчетов

Python

интерпретируемый универсальный язык программирования



R

интерпретируемый язык программирования для статистических вычислений



Fernando Pérez

Jupyter
2014

2007

Python 3
2008

2001



Travis Oliphant

Python
1991

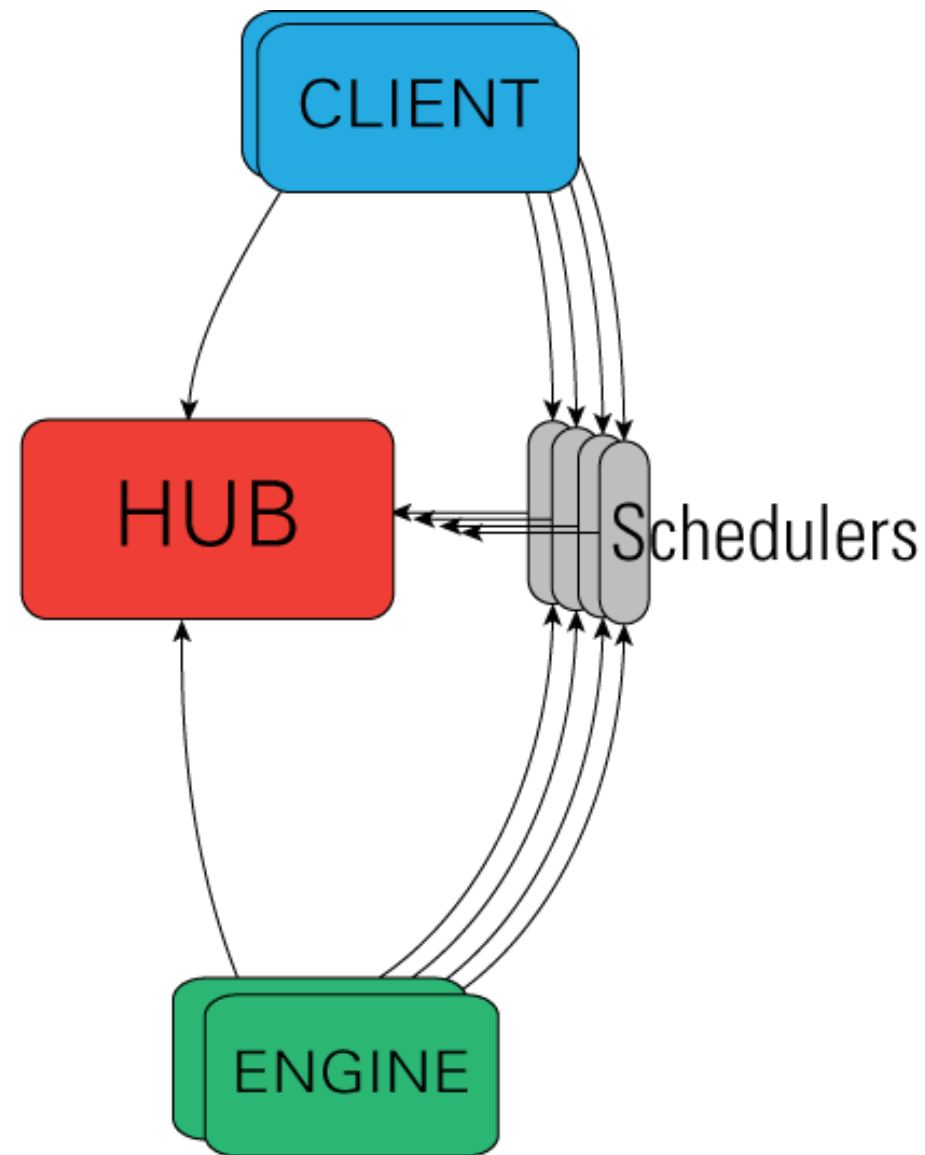
1993

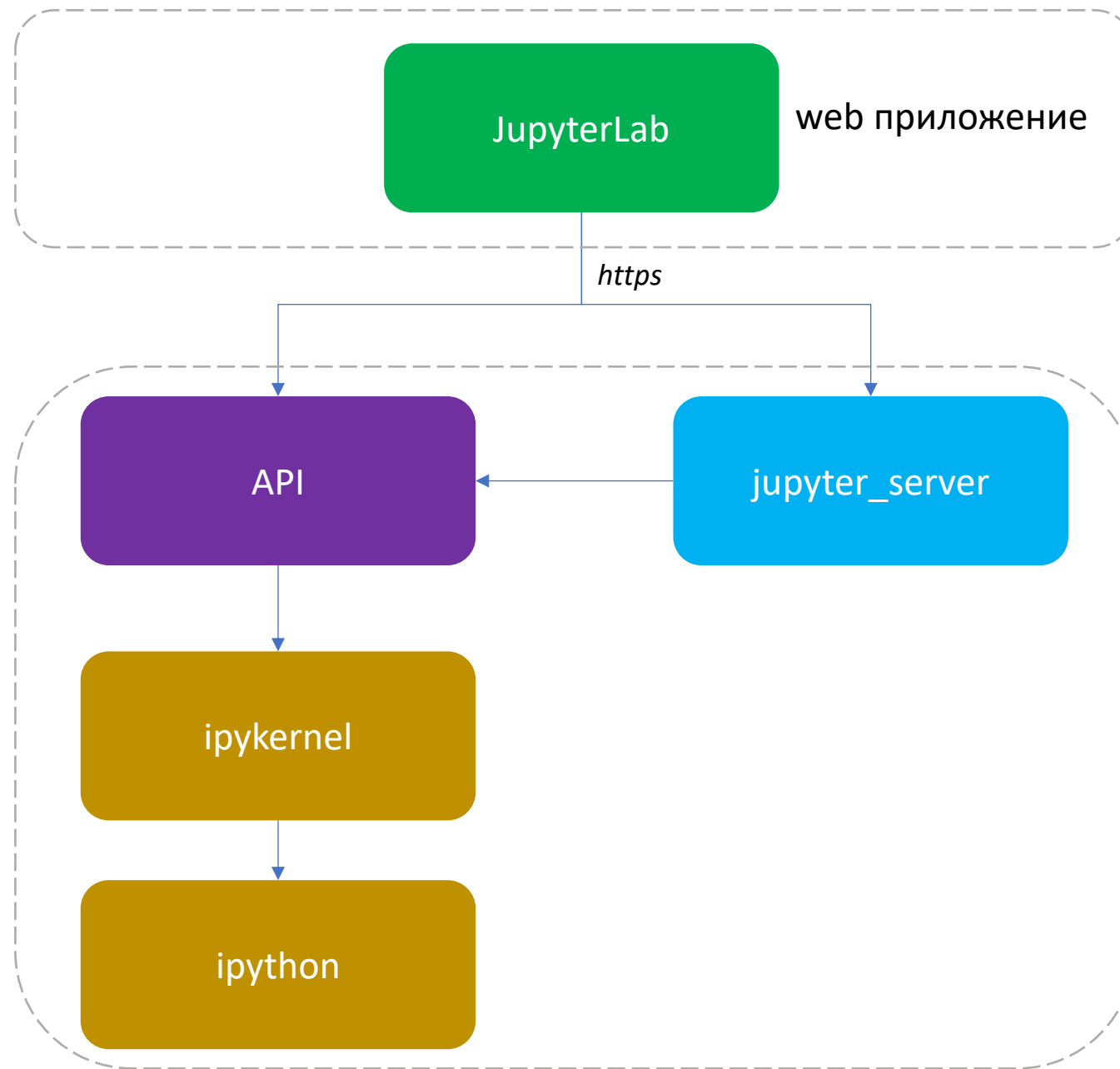


Guido van Rossum

Техническая архитектура

Architectural View of IPython's parallel machinery





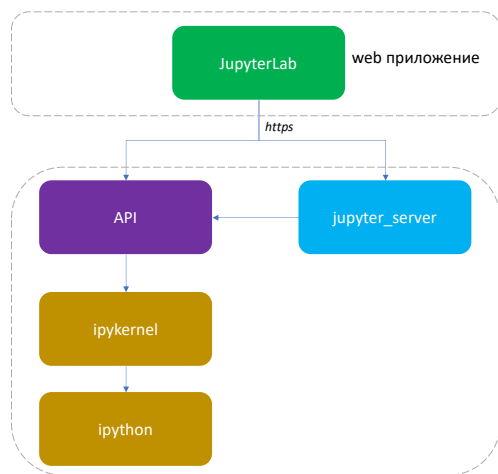
Другие ядра

IRKernel

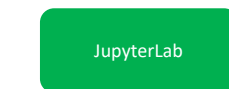
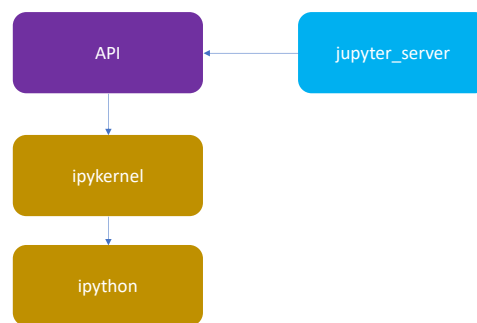
IJulia

Варианты использования

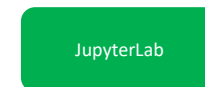
Все локально на компьютере разработчика



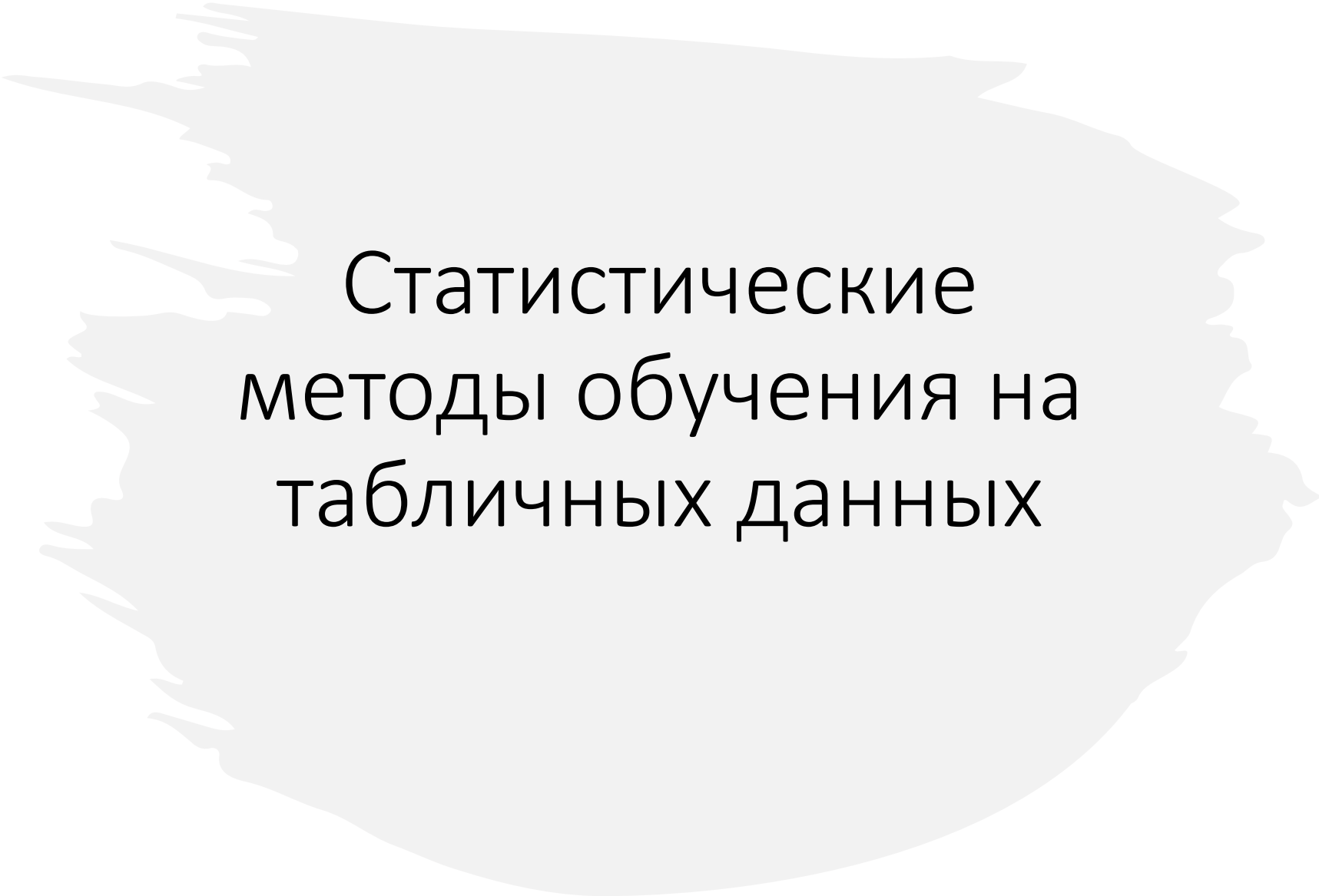
Клиент сервер



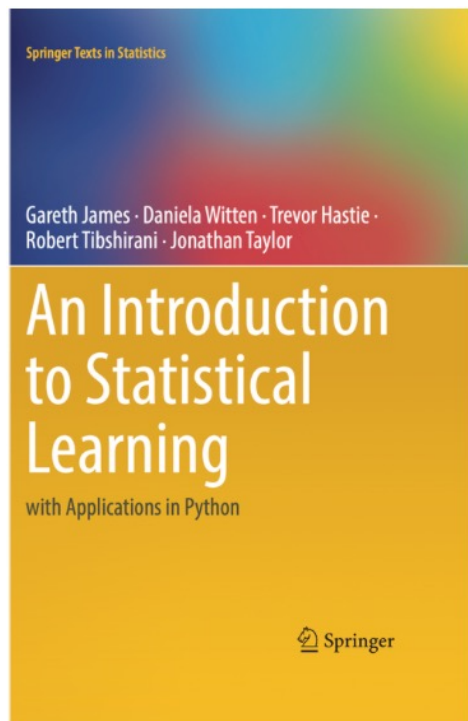
Внешняя инфраструктура



<https://colab.research.google.com>



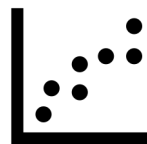
Статистические методы обучения на табличных данных



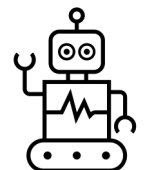
<https://www.statlearning.com>



Целью обучения является понимание и предвидение.






Статистический вывод (statistical inference) - это набор методов, которые позволяют формулировать суждения об общем (генеральная совокупность) на основании частного (выборка), оценивая меру уверенности в предсказании, вероятность ошибки.



Статистическая теория обучения имеет дело с задачами нахождения предсказательной функции на основе обучающего набора данных.

Формально

объекты	атрибуты		
		M	182
		F	165
		M	176

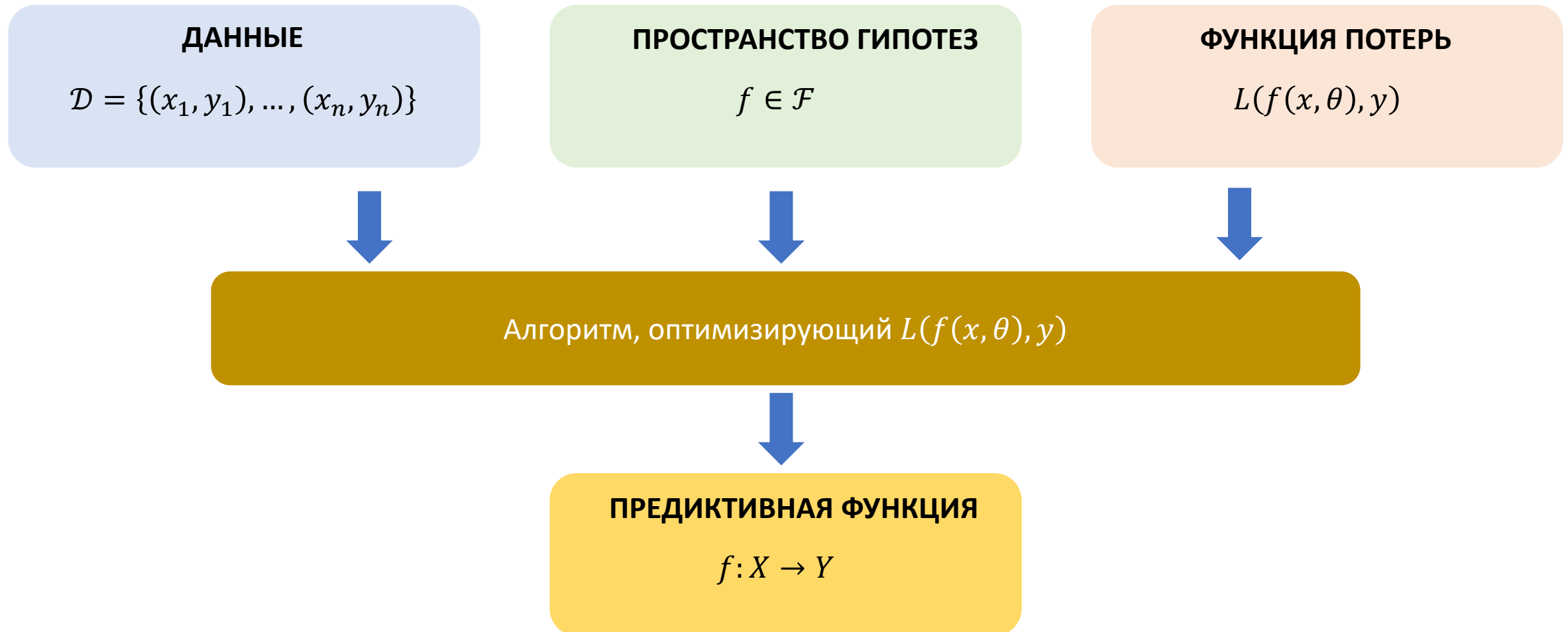
- Пусть X – векторное пространство *всех возможных* входных данных, а Y – векторное пространство *всех возможных* выходов.
- Статистическая теория обучения предполагает, что имеется некое *неизвестное* распределение вероятности над произведением пространств $Z = X \times Y$, т. е. существует некоторая неизвестная функция $p(z) = p(x, y)$.
- Обучающее множество данных состоит из n наблюдений, полученных из распределения $p(z)$:

$$\mathcal{D} = \{z_1, \dots, z_n\} = \{(x_1, y_1), \dots, (x_n, y_n)\}$$

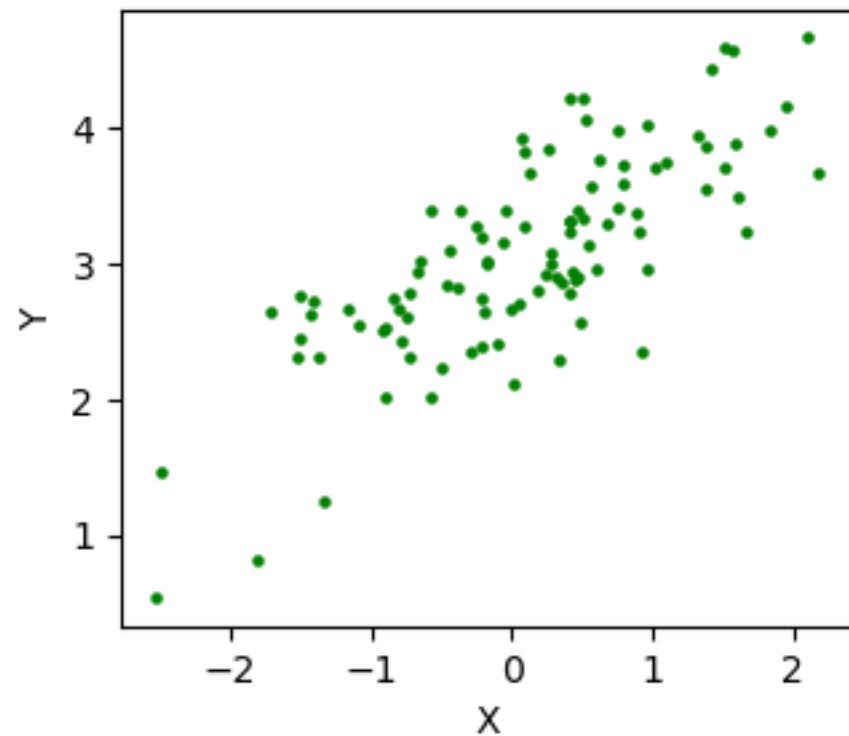
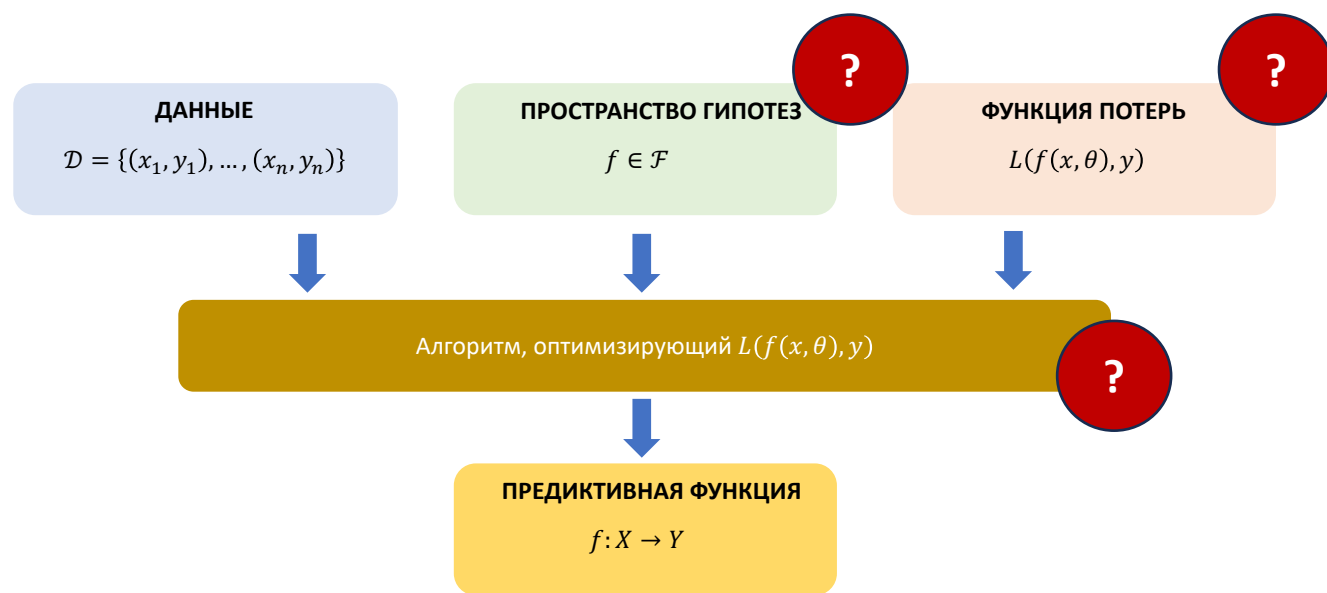
- Задача состоит в нахождении функции $f: X \rightarrow Y$ из функционального пространства \mathcal{F} , такой что $f(x, \theta) \sim y$. Здесь θ – вектор параметров.
- Пусть $L(f(x, \theta), y)$ - метрика (функция потерь), измеряющая разность между $f(x, \theta)$ и y .
- Тогда лучшая функция f это та, которая минимизирует *эмпирический риск*:

$$R(f) = \frac{1}{n} \sum_{i=1}^n L(f(x, \theta), y_i).$$

Метод статистического обучения



Линейная регрессия



$$\mathcal{F}: y \sim f(x) = a + bx + \varepsilon$$

$$\theta = \{a, b\}$$

Задача: найти наилучшие значения a и b , максимально приближающие значения $f(x)$ к y .

$$L(f(x, \theta), y) = (f(x, \theta) - y)^2$$

MSE

ДАННЫЕ

$$\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\}$$

ПРОСТРАНСТВО ГИПОТЕЗ

$$f \in \mathcal{F}$$

ФУНКЦИЯ ПОТЕРЬ

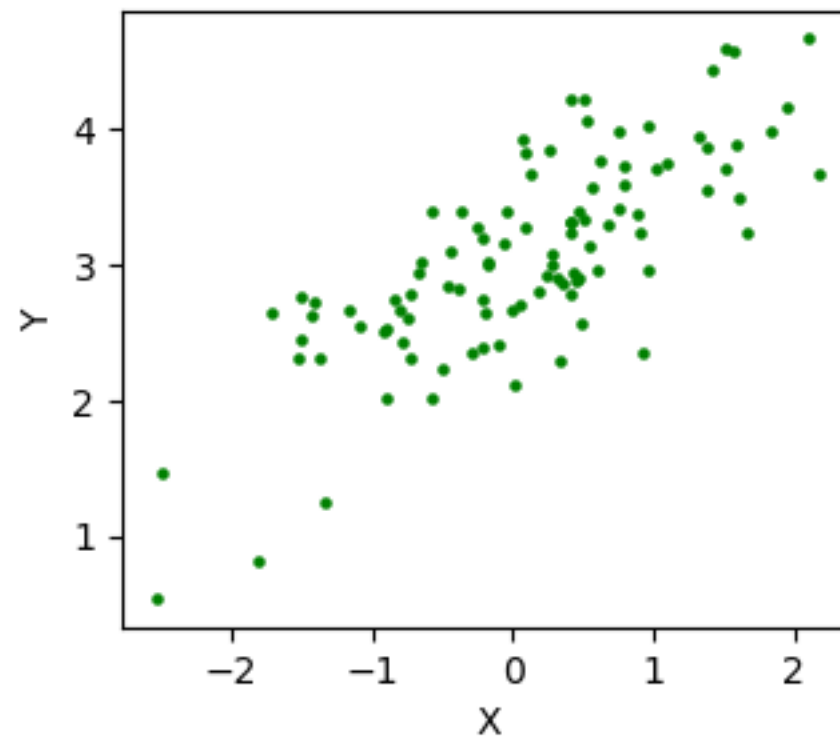
$$L(f(x), y)$$

Алгоритм, оптимизирующий $L(f(x), y)$

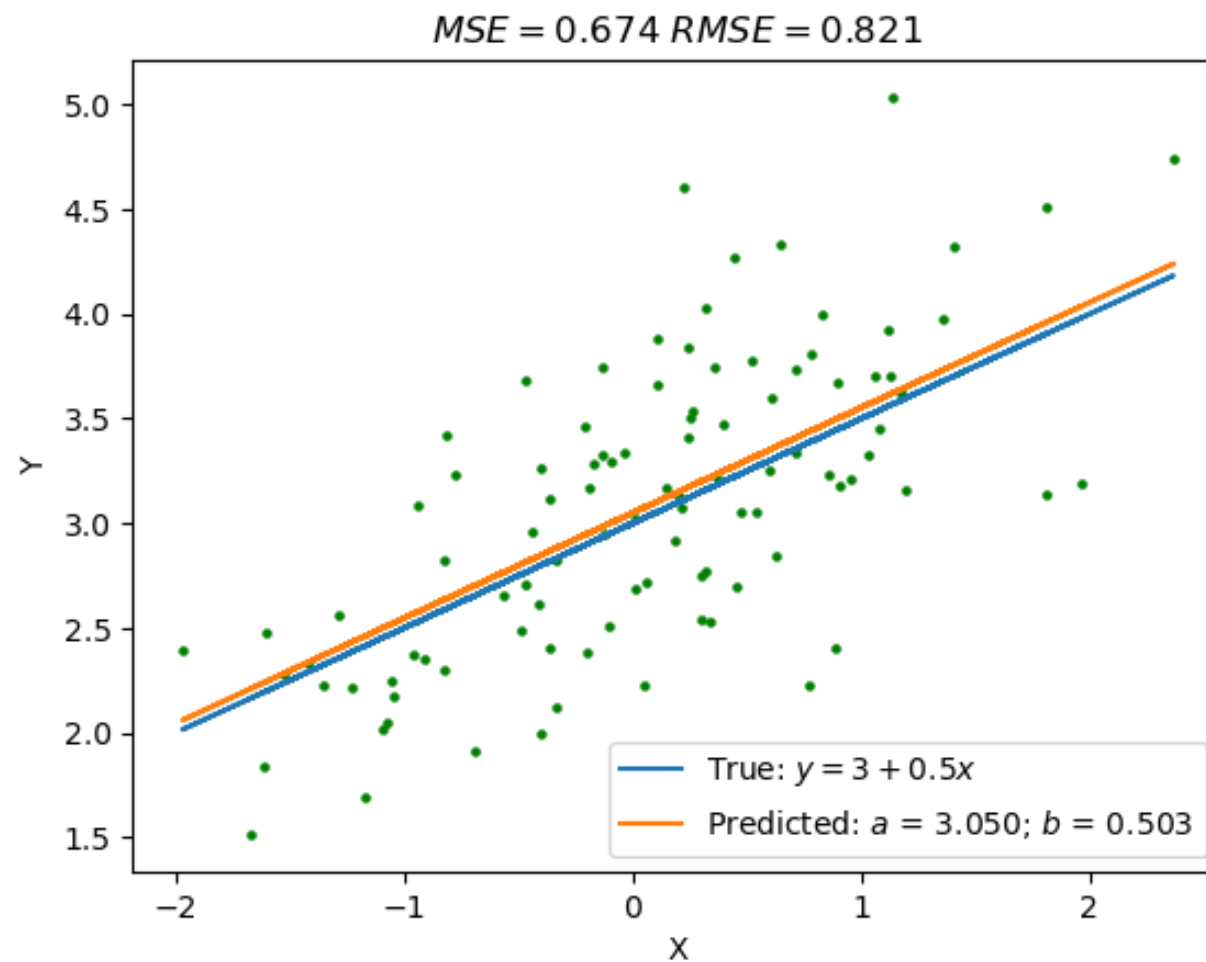
ПРЕДИКТИВНАЯ ФУНКЦИЯ

$$f: X \rightarrow Y$$

$$\frac{dL(f(x, \theta), y)}{d\theta} = 0$$



Результат



Базовый процесс ML

```
# Sklearn pipeline
```



```
lr = LinearRgeression()    # инициализируем объект, реализующий предиктивную модель  
lr.fit(X, y)               # обучаем модель  
y_pred = lr.predict(X)     # вычисляем предсказанные значения
```

Отличие от эконометрики

Эконометрика (статистика)

- Предполагается, что модель отражает каузальные связи в системе.
- Поэтому важно включить в модель все значимые факторы (и исключить незначимые).
- Величина коэффициентов регрессии имеет смысл – это изменение y при изменении x на единицу.
- Качество модели оценивается через ошибку на обучающих данных.

Машинное обучение

- Важные предиктивная (обобщающие) способность модели.
- Каузальные связи между факторами не интерпретируются.
- Больше данных – лучше модель.
- Качество модели оценивается как ошибка на данных, которые модель не видела.

Многомерная линейная регрессия

California Housing Dataset (20640 наблюдений, 8 атрибутов)

https://www.dcc.fc.up.pt/~ltorgo/Regression/cal_housing.html

Атрибуты:

MedInc медианный доход в квартале

HouseAge средний возраст дома в квартале

AveRooms среднее количество комнат на одно домохозяйство

AveBedrms среднее количество спален на одно домохозяйство

Population население квартала

AveOccup среднее количество членов домохозяйства

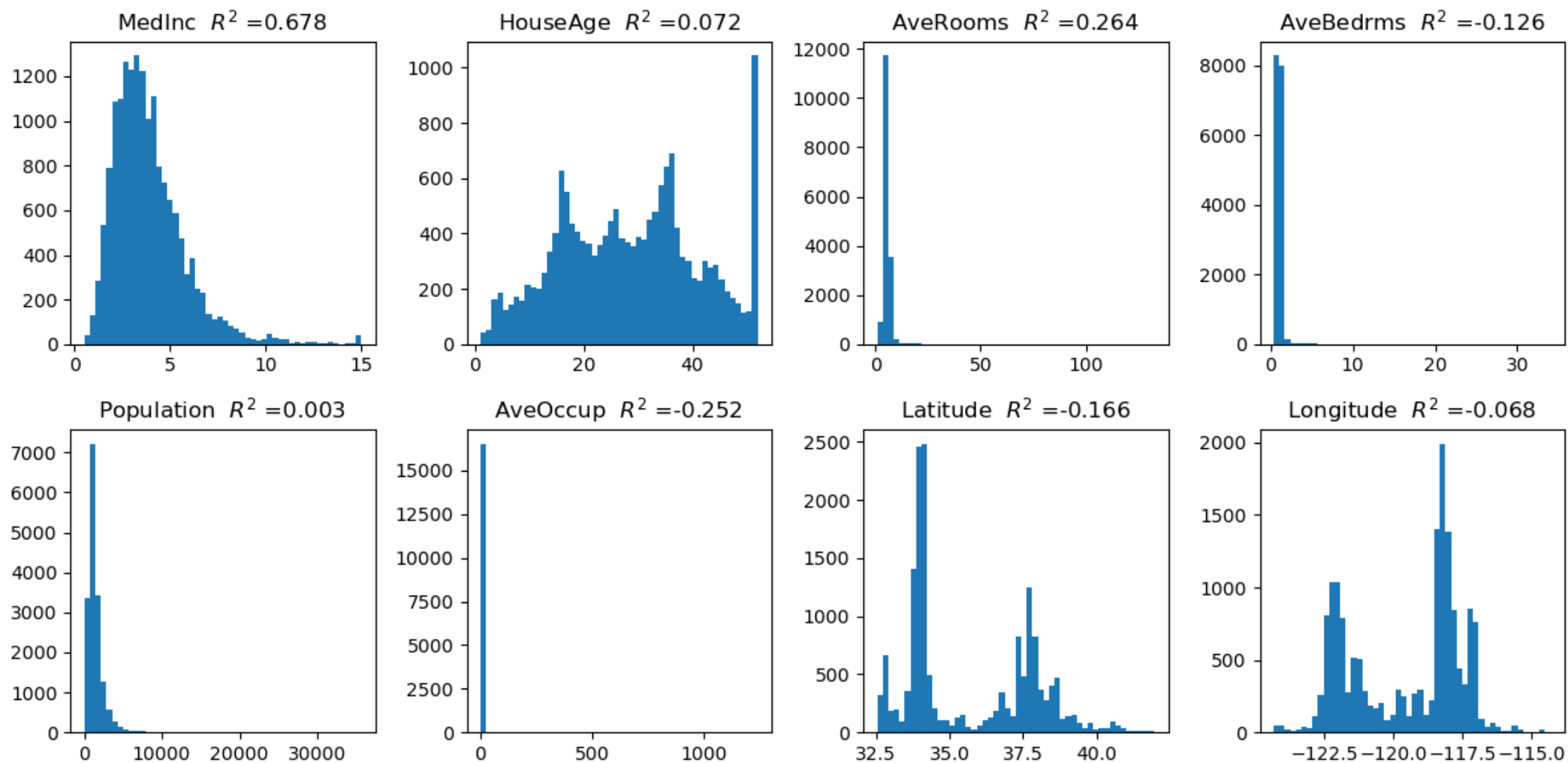
Latitude широта, на которой расположен квартал

Longitude долгота, на которой расположен квартал

Целевая переменная:

MedHouseVal медианная стоимость дома в квартале в сотнях тысяч долларов (\$100 000)

Посмотрим на данные

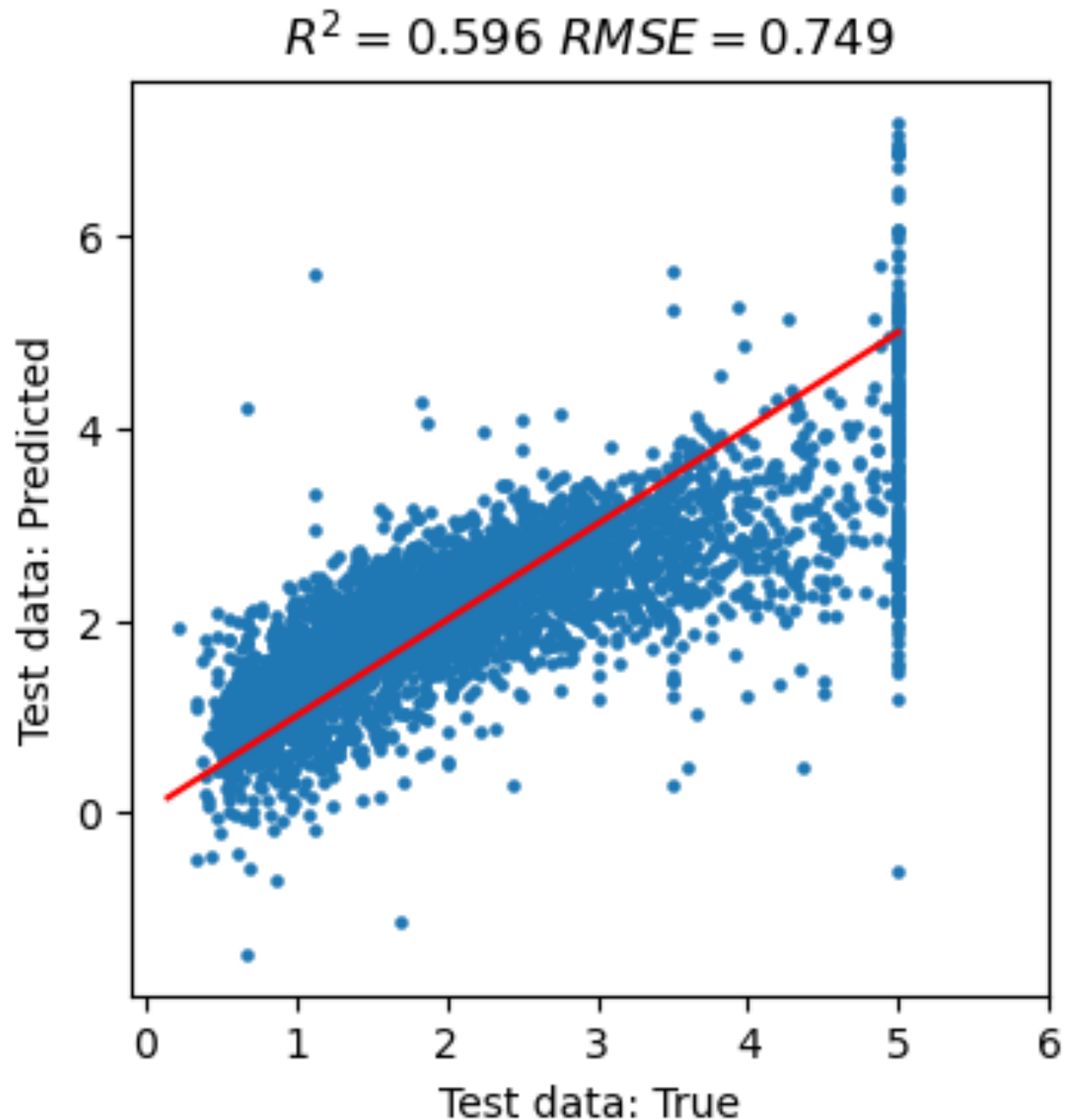



Результат

Коэффициенты

Intercept : -35.497
MedInc : 0.444
HouseAge : 0.009
AveRooms : -0.120
AveBedrms : 0.638
Population: -0.000
AveOccup : -0.005
Latitude : -0.407
Longitude : -0.419

В ML эти значения
интерпретируемого смысла
не имеют!





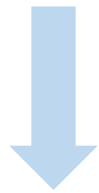
Классификация

Регрессия и классификация

$$\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\}$$

X – внешние переменные

Y – целевая переменная (ее надо предсказать)



$$Y \in \mathbb{R}$$

y – вещественное число

Регрессия



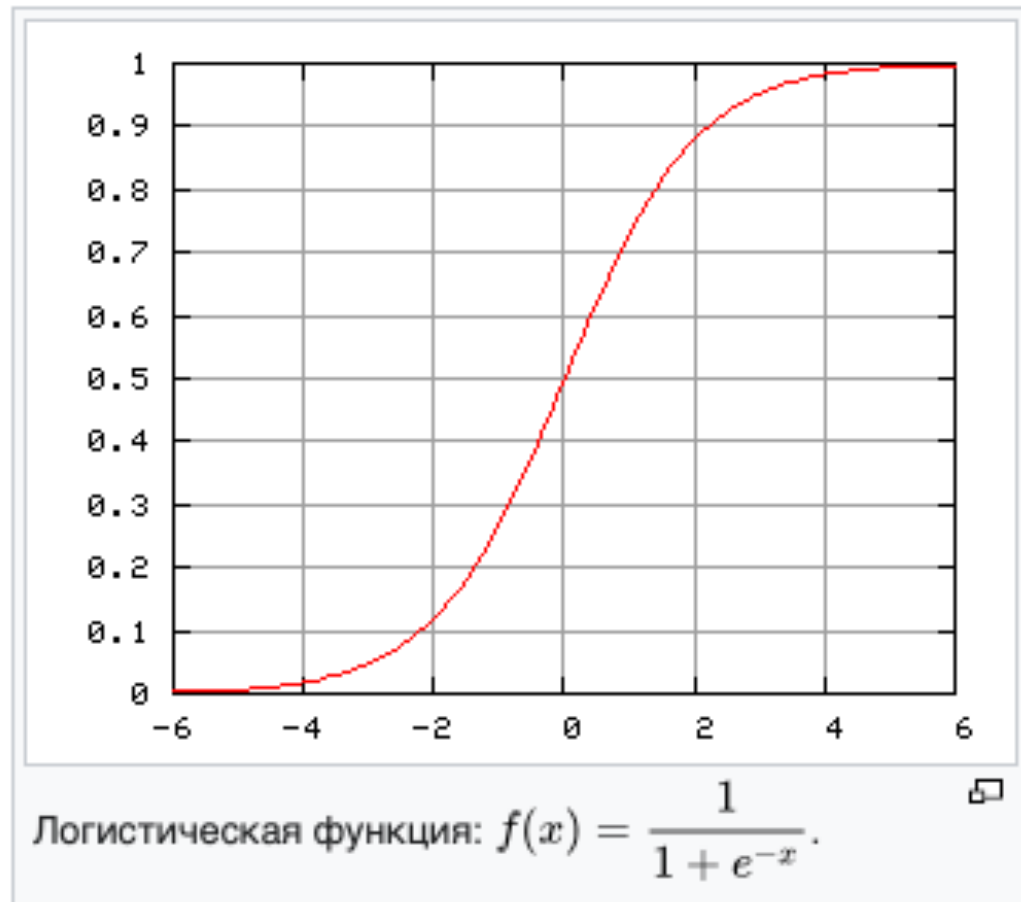
$$Y \in \{0, 1, \dots, m\}$$

y принимает значения из конечного множества

Классификация

Чаще всего рассматриваются 2 класса, т.е.
 $Y \in \{0, 1\}$ – бинарная классификация

Модель: логистическая регрессия



Логистическая регрессия (logit model) — статистическая модель, используемая для прогнозирования вероятности возникновения некоторого события путём его сравнения с логистической кривой.

$Y \in \{0,1\}$ - зависимая переменная, принимающая лишь одно из двух значений: 0 (событие не произошло) и 1 (событие произошло),

$X \in \mathbb{R}^k$ - множество вещественных независимых переменных (признаков, предикторов, регрессоров), на основе значений которых требуется вычислить вероятность принятия того или иного значения зависимой переменной.

$$f(z) = \frac{1}{1 + e^{-z}}$$

$$z = \theta_0 + \theta_1 x_1 + \dots + \theta_k x_k$$

Функция потерь - 1

Вероятность позволяет нам предсказывать неизвестные результаты, основанные на известных параметрах: $P(x) = P(x|\theta)$

Правдоподобие позволяет нам оценивать неизвестные параметры, основанные на известных результатах.: $L(\theta) = L(\theta|x = X)$



Функция правдоподобия — это совместное распределение выборки из параметрического распределения, рассматриваемое как функция параметра.

Она позволяет оценить несколько распределений с разными параметрами и оценить для какого из них наблюдаемые значения наиболее вероятны.

Найти параметры θ , максимизирующие значение функции правдоподобия

$$L(\theta) = \prod_{i=1}^n P(y = y_i | x = x_i)$$

на обучающей выборке

$$\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\}$$

Функция потерь - 2

Задача оптимизации:

$$\hat{\theta} = \operatorname{argmax}_{\theta} L(\theta) = \operatorname{argmax}_{\theta} \prod_{i=1}^n P(y = y_i | x = x_i)$$

Максимизация функции эквивалента максимизации ее логарифма:

$$\ln L(\theta) = \sum_{i=1}^n \ln P(y = y_i | x = x_i) = \sum_{i=1}^n [y_i \ln f(x_i, \theta) + (1 - y_i) \ln(1 - f(x_i, \theta))]$$

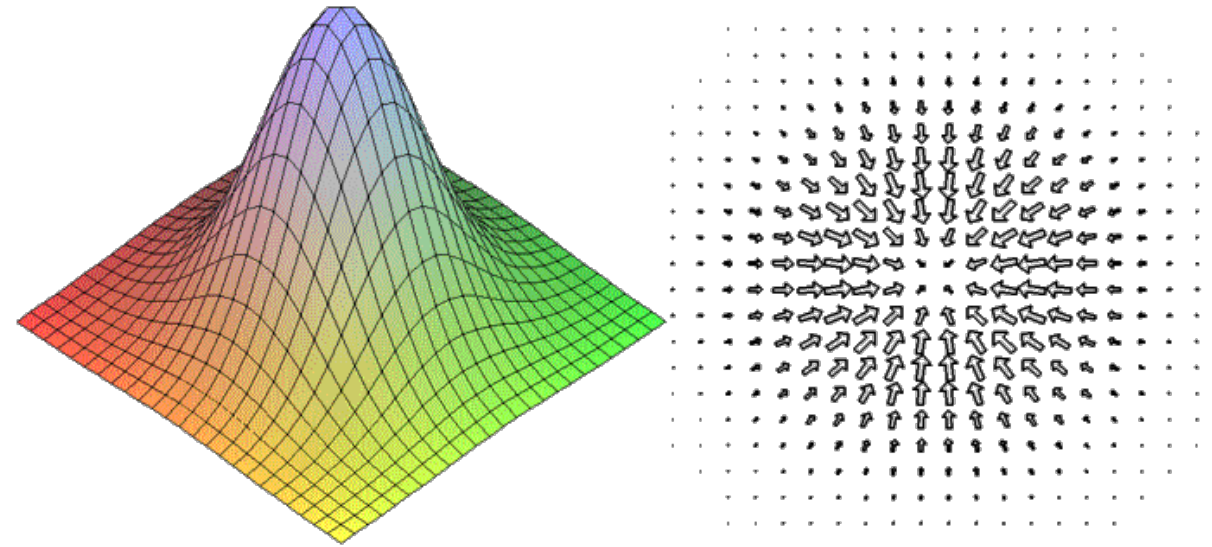
Метод оптимизации: градиентный спуск

Градиент — вектор, своим направлением указывающий направление наискорейшего роста некоторой скалярной величины f , значение которой меняется от одной точки пространства к другой, образуя скалярное поле.

По модулю (величине) ∇f равен скорости роста величины f в направлении вектора.

Если f функция k переменных x_1, \dots, x_k , то ее градиентом называется k -мерный вектор

$$\left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_k} \right)$$



Основная идея метода градиентного спуска (при минимизации) заключается в том, чтобы идти в направлении наискорейшего спуска, а это направление задаётся антиградиентом $-\nabla f$

$$\theta^{[j+1]} = \theta^{[j]} - \lambda^{[j]} \nabla f(x, \theta), \quad \lambda - \text{скорость градиентного спуска}$$

$$f(z) = \frac{1}{1 + e^{-z}}$$

$$z = \theta_0 + \theta_1 x_1 + \dots + \theta_k x_k$$

$$L(f(x, \theta), y) = \sum_{i=1}^n [y_i \ln f(x_i, \theta) + (1 - y_i) \ln(1 - f(x_i, \theta))]$$

ДААННЫЕ

$$\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\}$$

ПРОСТРАНСТВО ГИПОТЕЗ

$$f \in \mathcal{F}$$

ФУНКЦИЯ ПОТЕРЬ

$$L(f(x), y)$$

Алгоритм, оптимизирующий $L(f(x), y)$

ПРЕДИКТИВНАЯ ФУНКЦИЯ

$$f: X \rightarrow Y$$

Задача
классификации

Градиентный спуск

$$\theta^{[j+1]} = \theta^{[j]} + \lambda^{[j]} \nabla L(f(x, \theta), y)$$

Пример: предсказание банкротств

Данные: 2457 компаний, из них 456 банкротов, 12 атрибутов

Финансовые коэффициенты					
	0	1	2	3	4
Коэффициент текущей ликвидности	1,855	2,195	0,546	2,075	0,328
Коэффициент быстрой ликвидности	0,495	1,760	0,501	1,739	0,203
Коэффициент абсолютной ликвидности	0,119	0,255	0,004	0,061	0,007
Коэффициент оборота задолженности	7,863	3,939	2,401	2,934	3,588
Коэффициент оборачиваемости средств	2,379	2,902	2,094	2,361	1,706
Соотношение собственных и заемных средств	10,701	1,073	1,624	1,506	-2,512
Коэффициент собственных оборотных средств	-0,393	-0,820	-0,956	0,310	-2,075
Коэффициент автономии	0,085	0,482	0,378	0,399	-0,622
Степень платежеспособности по текущим обязательствам	2,711	1,945	15,555	2,584	21,479
Рентабельность чистых активов по чистой прибыли	0,114	0,130	0,283	0,247	0,888
Рентабельность	0,158	0,167	0,286	0,308	1,030
Рентабельность продаж	0,040	0,165	0,253	0,051	0,119
Банкрот	0	0	0	0	0

Построение классификатора

```
logr = LogisticRegression(solver = 'liblinear', random_state = 1)
```

```
logr.fit(X_train, y_train)
```

```
y_pred_train = logr.predict(X_train)
```

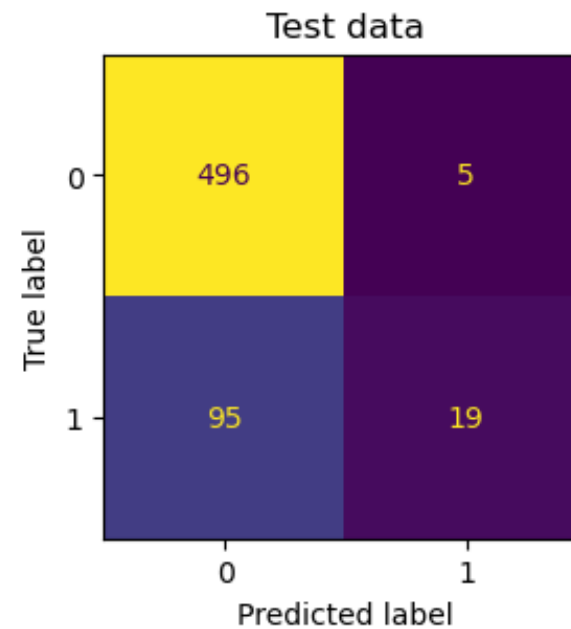
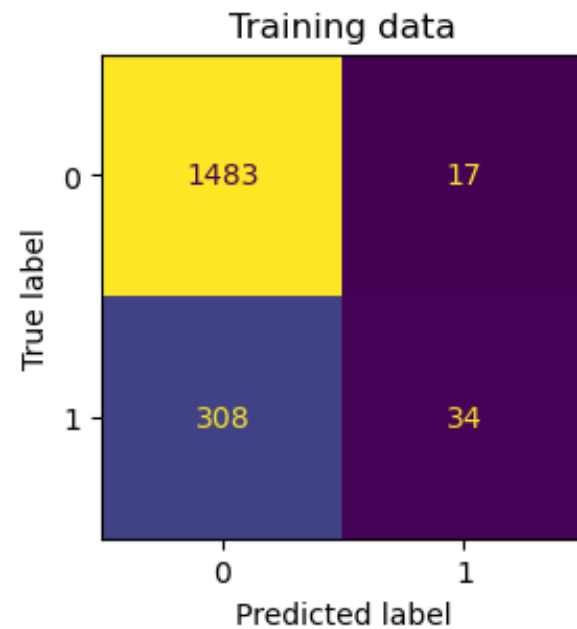
```
y_pred_test  = logr.predict(X_test)
```

Confusion Matrix / Матрица ошибок

		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$

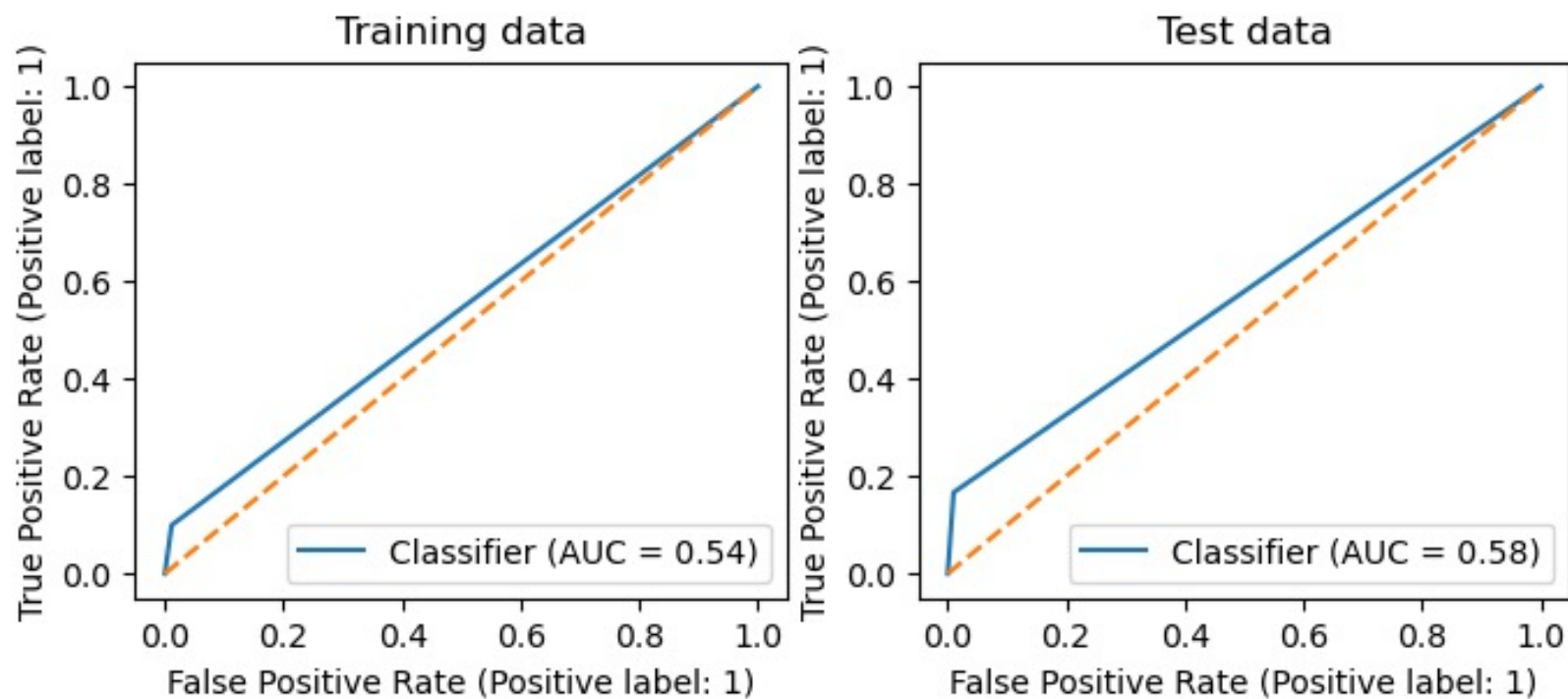
Accuracy: Train 0.824 Test 0.837

Результат: confusion matrix



Accuracy: Train 0.824 Test 0.837

Результат: ROC AUC



Обучение без учителя




Кластерный анализ



Статистическое (машинное) обучение






Обучение с учителем
Supervised Learning

объекты	атрибуты		метки
		182 89	
		165 58	
		176 68	



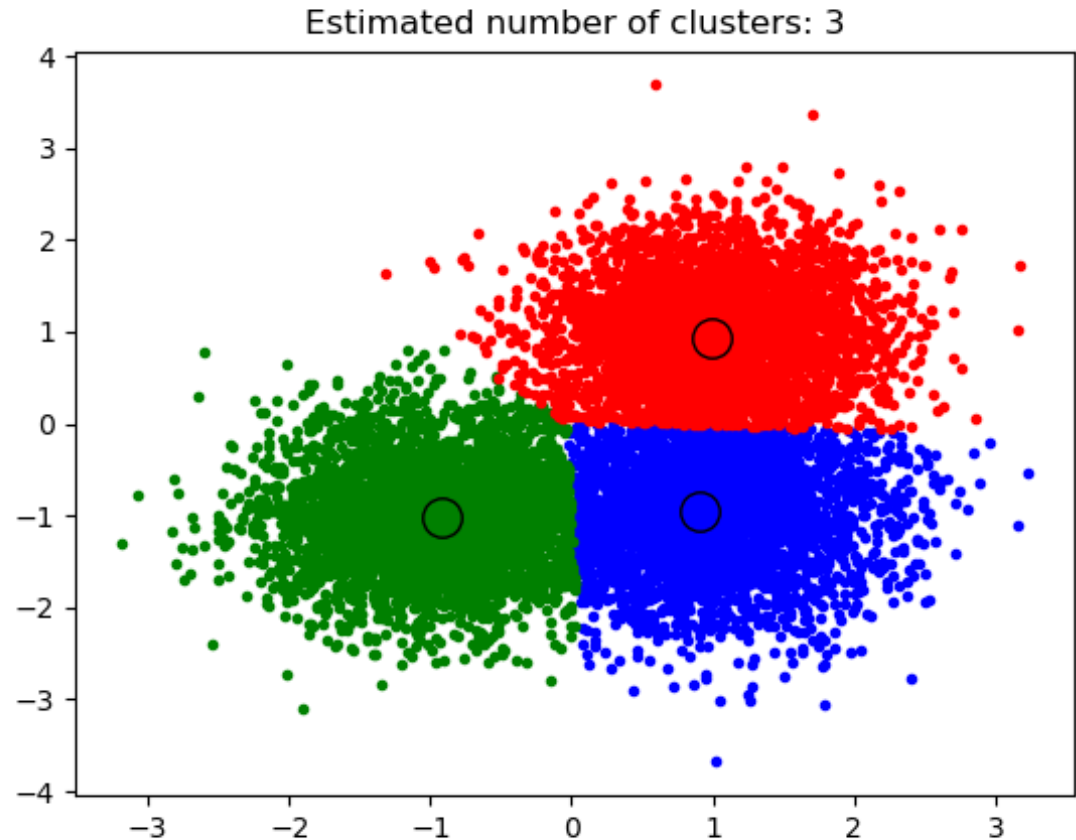
Обучение без учителя
Unsupervised Learning

объекты	атрибуты	
		182 89
		165 58
		176 68

Обучение без учителя. Кластеризация

Примеры:

- На какое количество классов делятся клиенты в зависимости от их покупок?
- Какие группы можно выделить среди предприятий региона?
- Как можно разделить сотрудников по их компетенциям?



Source: <http://scikit-learn.org/stable/modules/clustering.html>

$$C = f(x, \theta)$$

C – метка кластера, θ – параметры функции

$$C = \min_{\mu_j \in C} \|x_i - \mu_j\|$$

ДАННЫЕ

$$\mathcal{D} = \{x_1, \dots, x_n\}$$

ПРОСТРАНСТВО ГИПОТЕЗ

$$f \in \mathcal{F}$$

ФУНКЦИЯ ПОТЕРЬ

$$L(f(x, \theta), y)$$

Кластерный анализ

Алгоритм, оптимизирующий $L(f(x, \theta), y)$

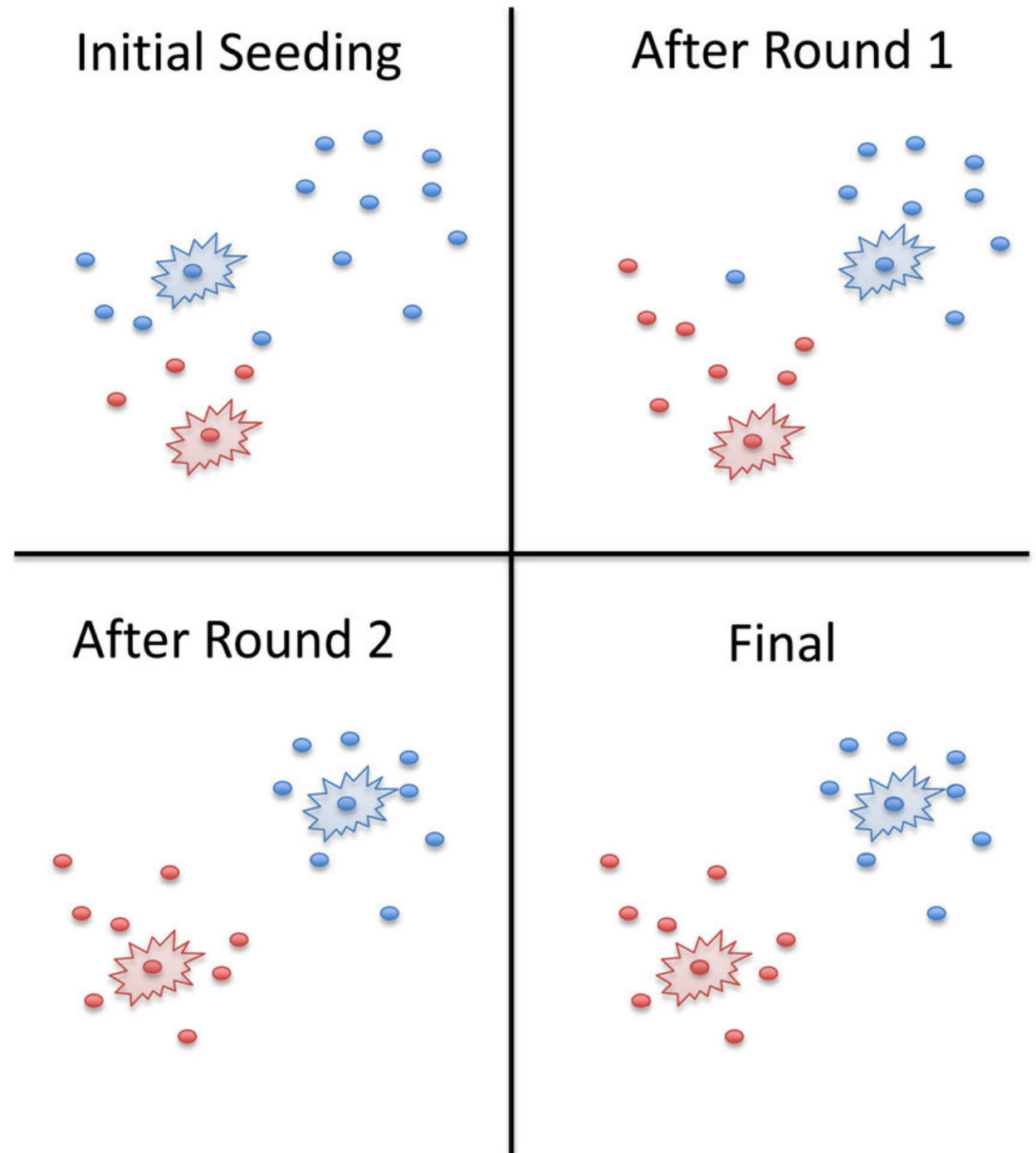
ПРЕДИКТИВНАЯ ФУНКЦИЯ

$$f: X \rightarrow C$$

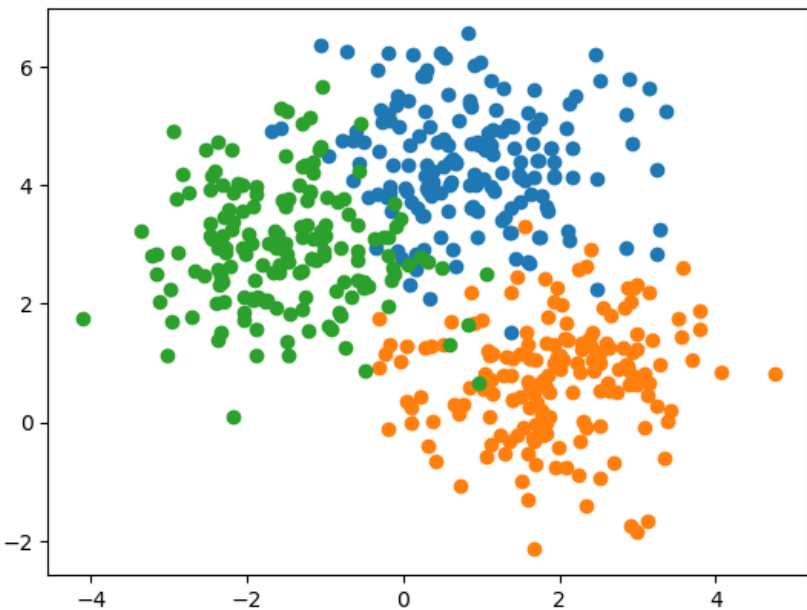
k -means

разбивает набор данных из n наблюдений на k непересекающихся кластеров C , каждый из которых описывается средним значением μ_j объектов в кластере

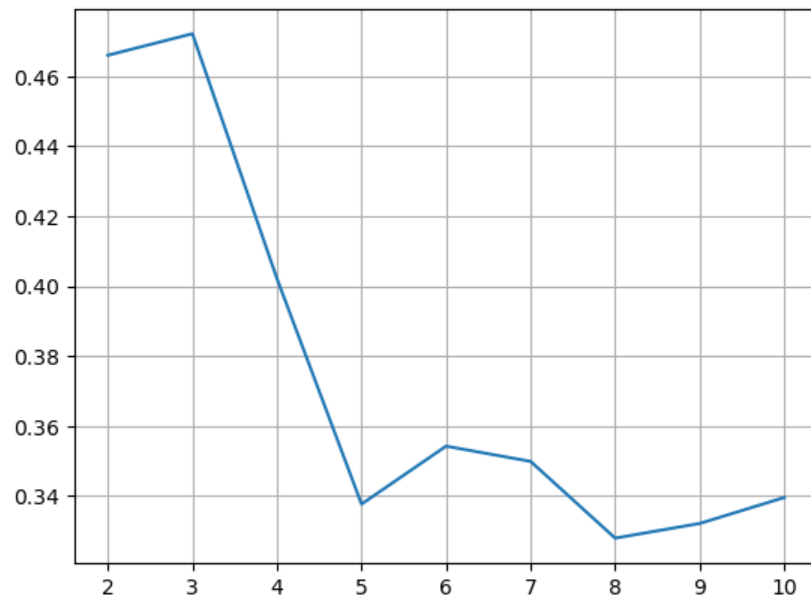
Алгоритм k-means



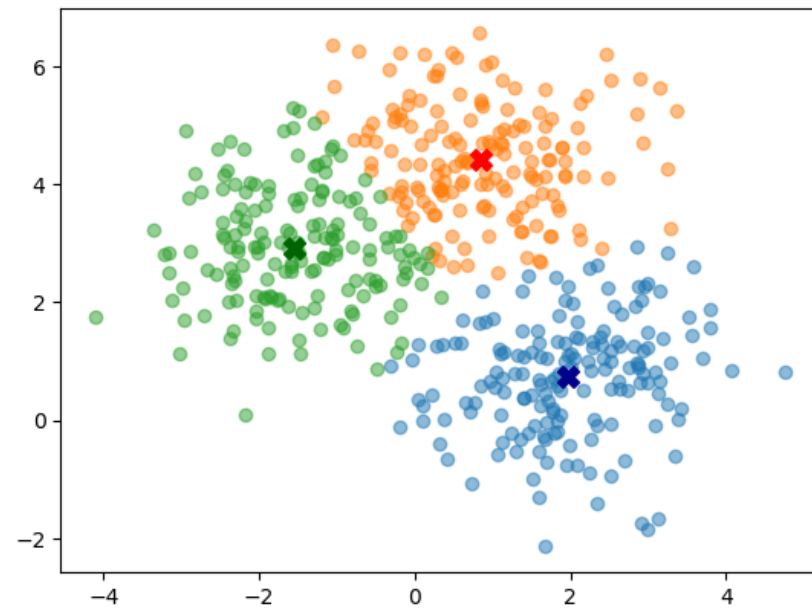
Простой пример



(1) Набор данных

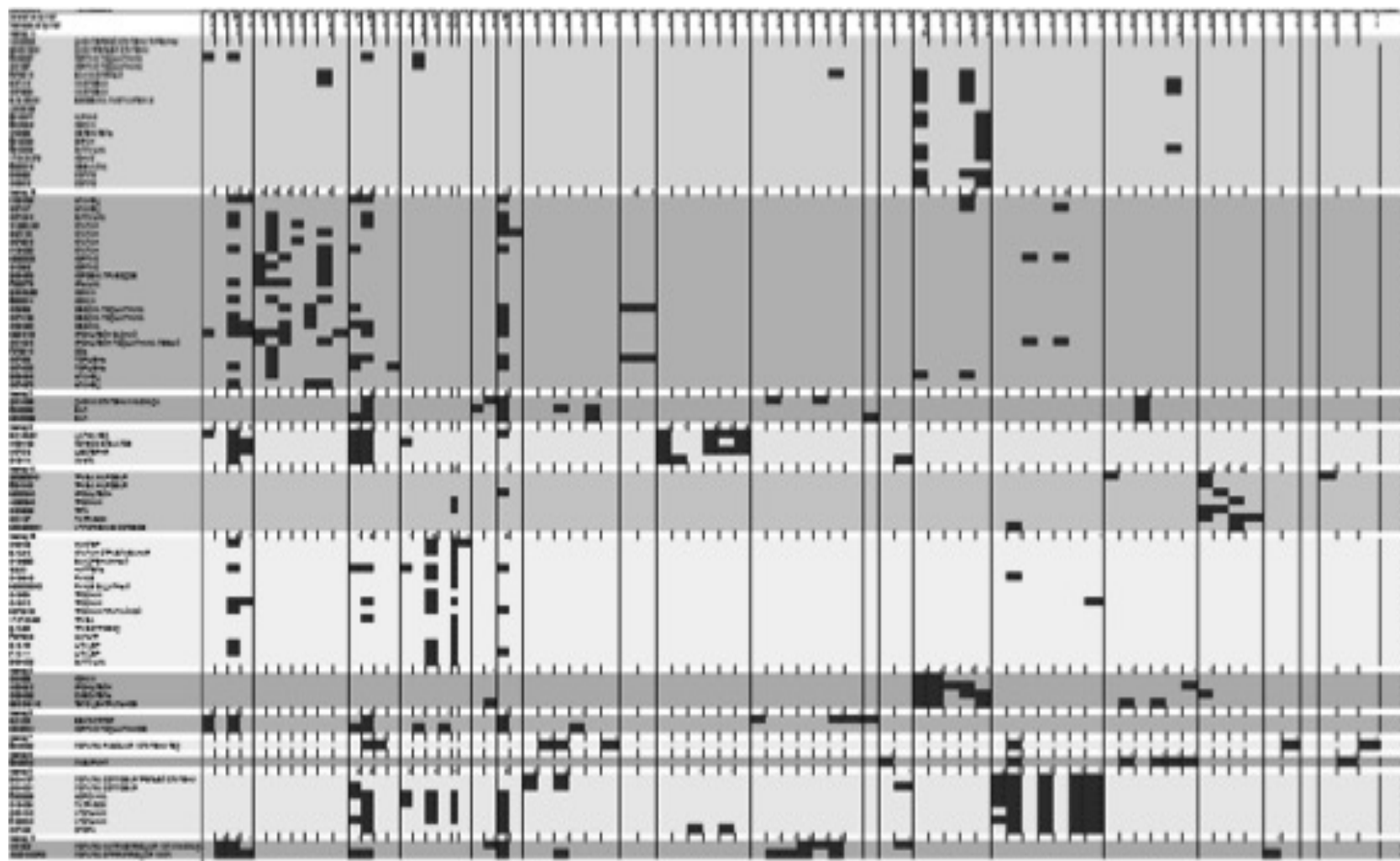


(2) Определяем число кластеров, используя метрику silhouette



(3) Результат

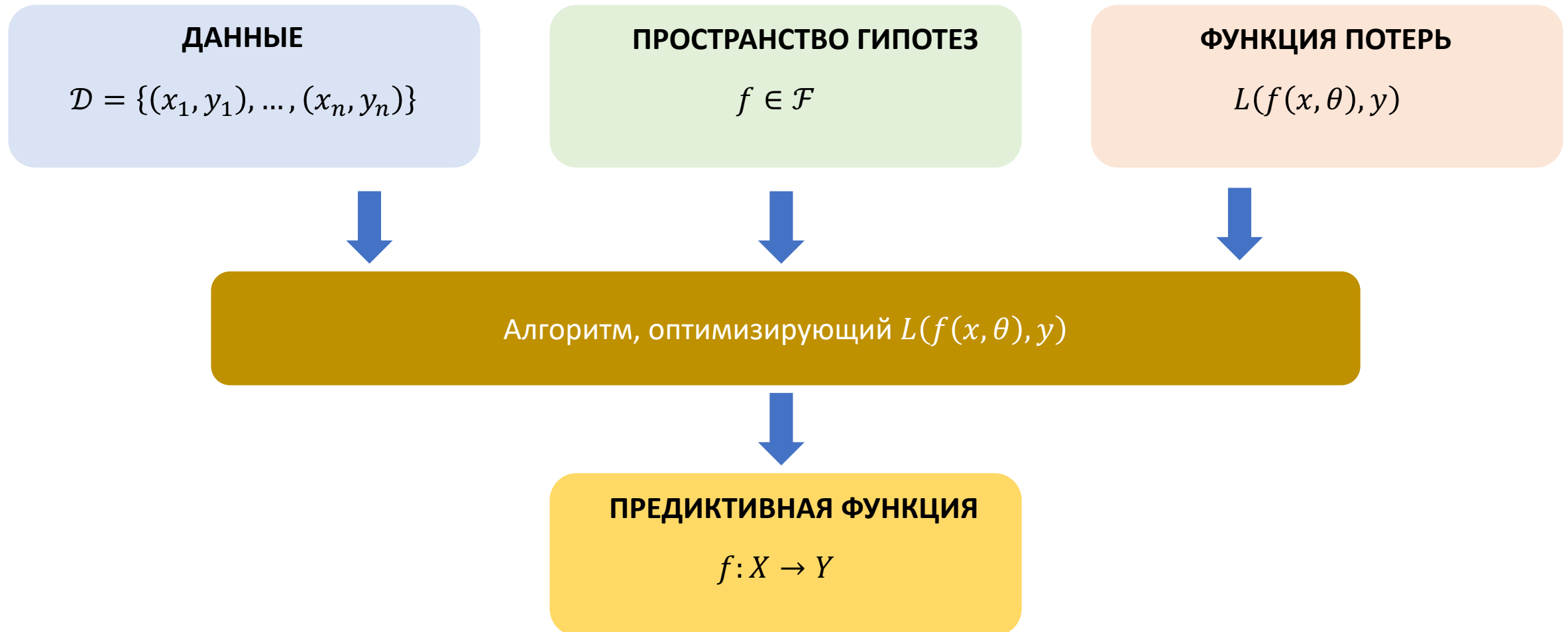
Реальный кейс



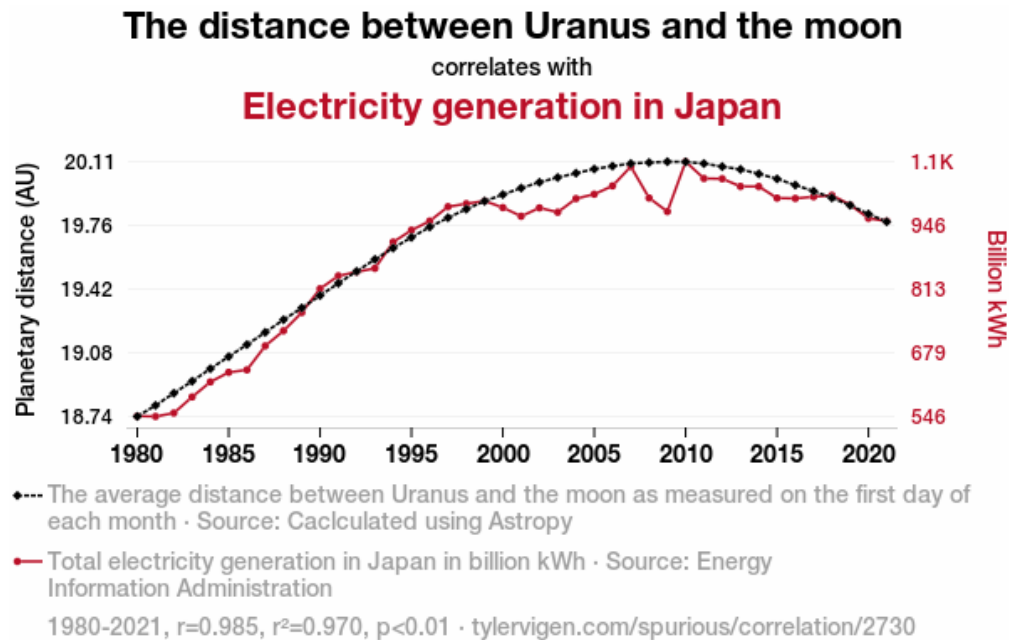


Ресар: Что мы узнали?

Общая формулировка задачи ML



Ограничения статистического обучения



<https://tylervigen.com>

- Предиктивная функция строится за счет статистически значимых ассоциаций между переменными.
- Наличие ассоциации между двумя переменными не означает, что между ними существует причинно-следственная связь!
- Поэтому интерпретация моделей ML (извлечение знаний) чаще всего невозможна или затруднена.



Вопросы?