SCHOOL
OF ENERGY
& POWER ENGINEERING

Введение и основные понятия машинного обучения

Сергей Владимирович Аксёнов,

Доцент отделения информационных технологий ИШИТР,

Томский политехнический университет

Томск-2023

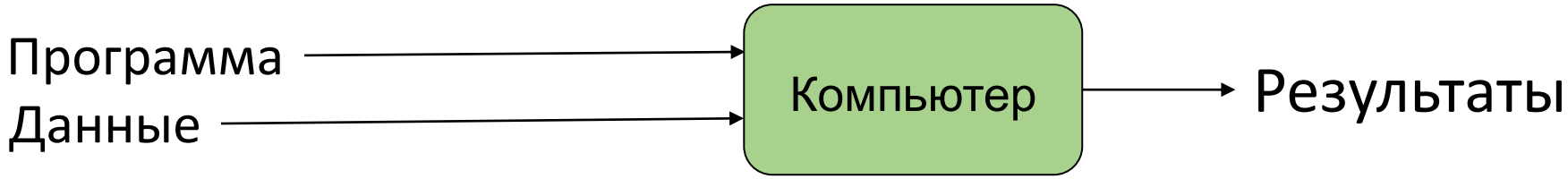
Определение

Машинное обучение — это программирование компьютеров для оптимизации критерия качества решения задач анализа данных с использованием примеров данных или прошлого опыта.

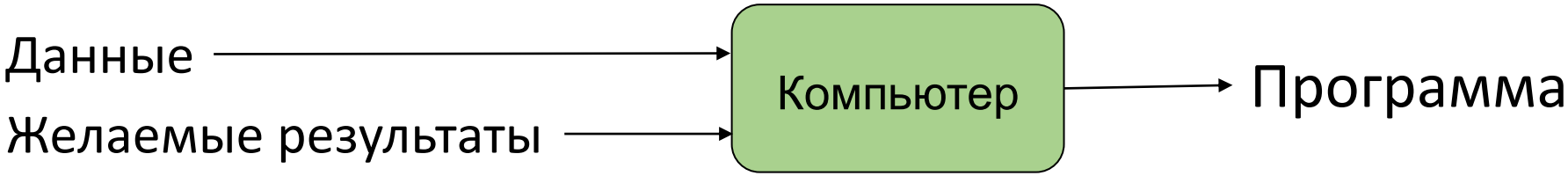
- ✓ Разработка приложений, которые трудно или дорого проектировать вручную, т.к. они требуют специфических знаний, навыков и опыта, связанных с исследуемой задачей
- ✓ Разработка систем, способных адаптироваться и настраиваться под конкретного пользователя
- ✓ Нахождение новых знаний в больших базах данных

Машинное обучение VS Традиционное программирование

Традиционное программирование



Машинное обучение

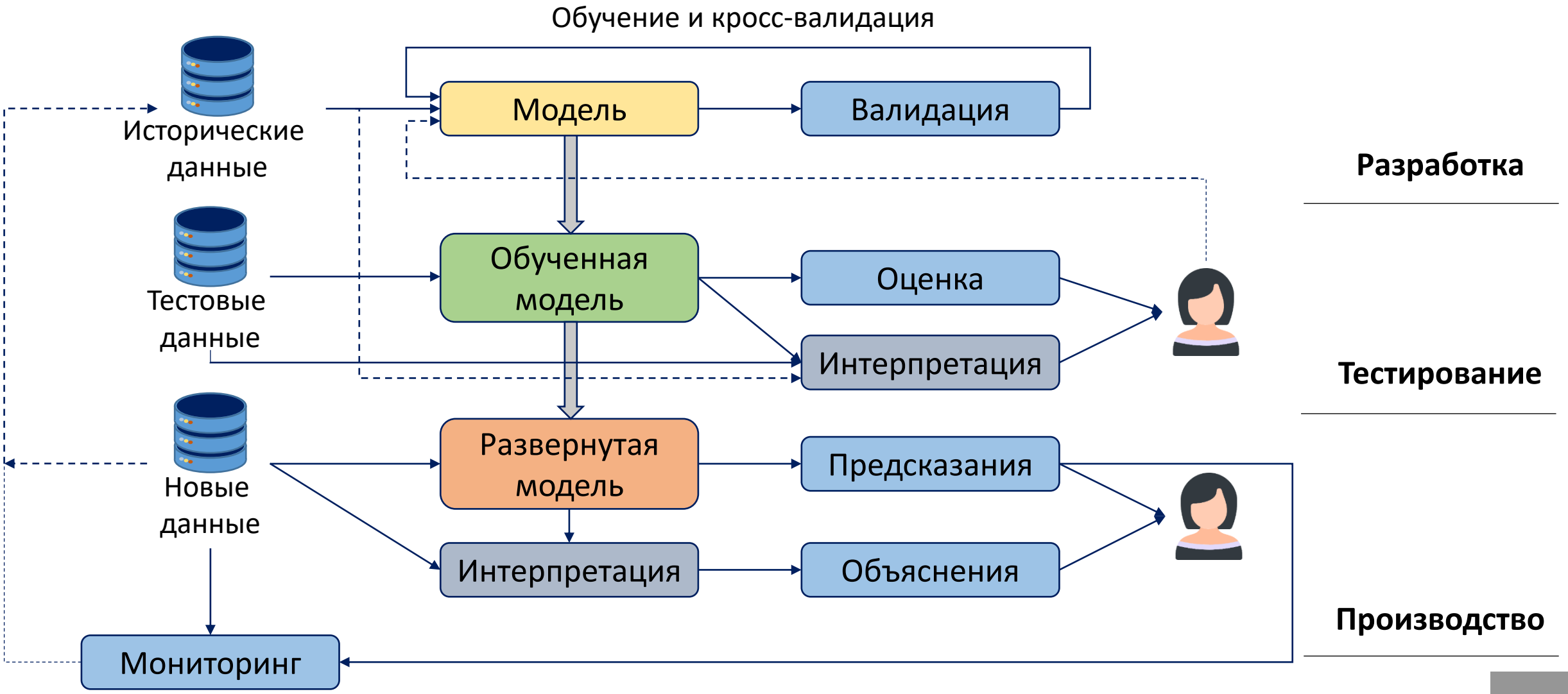


Основные задачи, решаемые машинным обучением

- Классификация
- Регрессия
- Кластеризация
- Сокращение размерности
- Ассоциации

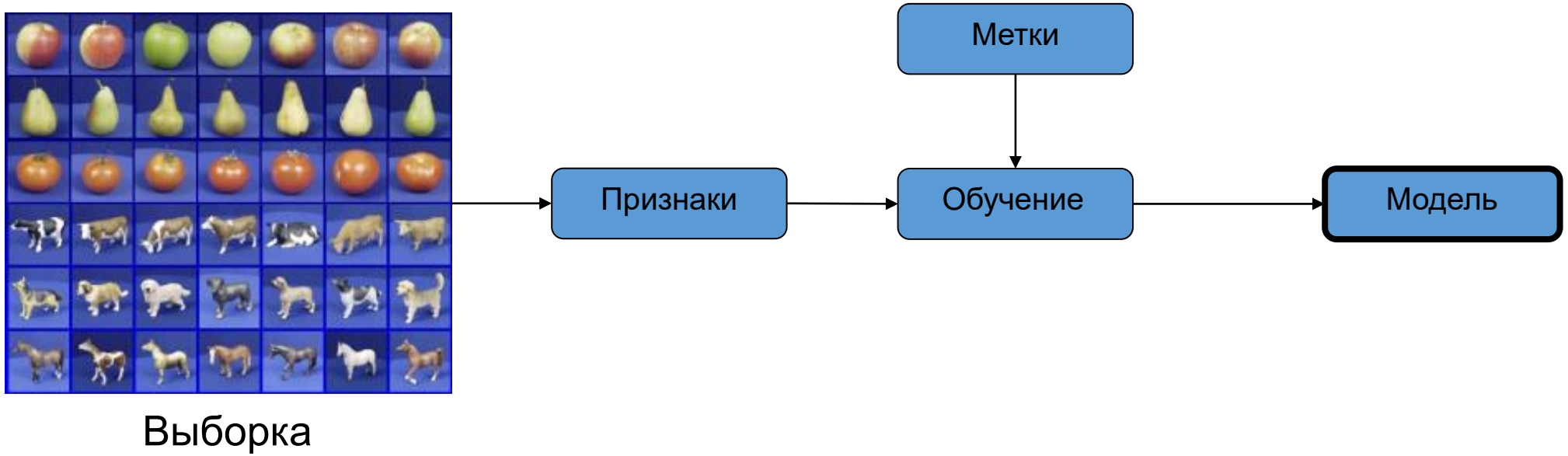
+ Некоторые дополнительные алгоритмы, повышающие эффективность работы : отбор признаков, обнаружение выбросов и т.д.

Процесс разработки надежных моделей

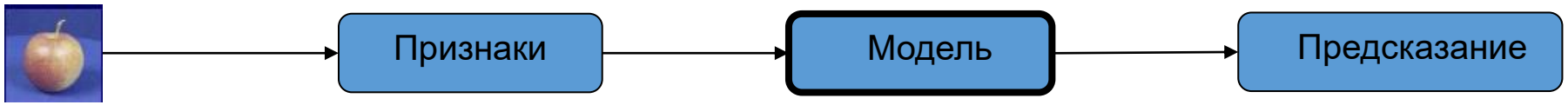


Обучение с учителем

Проектирование модели



Тестирование / Использование



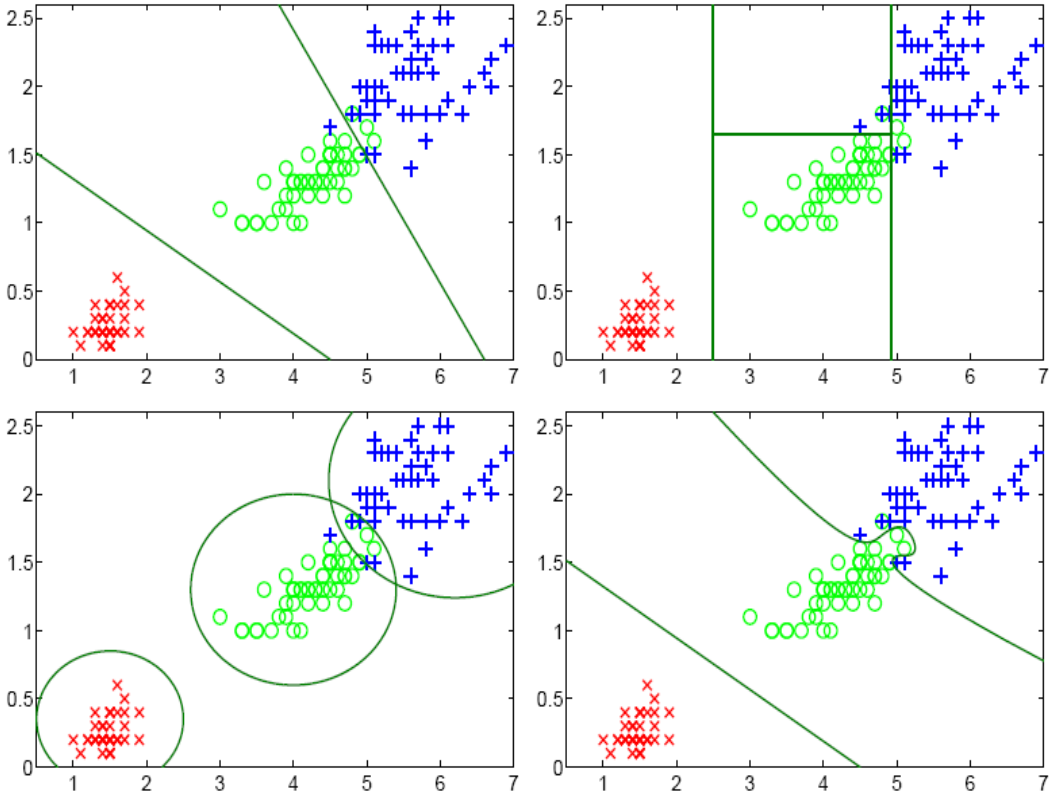
Классификация. Пример

Цель обучения: разделить пространство признаков на регионы, в которых располагаются объекты принадлежащие только одному классу.

Разделение пространства признаков (длина и ширина лепестка ириса) разными алгоритмами

Линейная модель

Дерево решений



Гауссовы смеси

Метод опорных векторов

Классы объектов

setosa

virginica

versicolor

Оценка качества моделей классификации

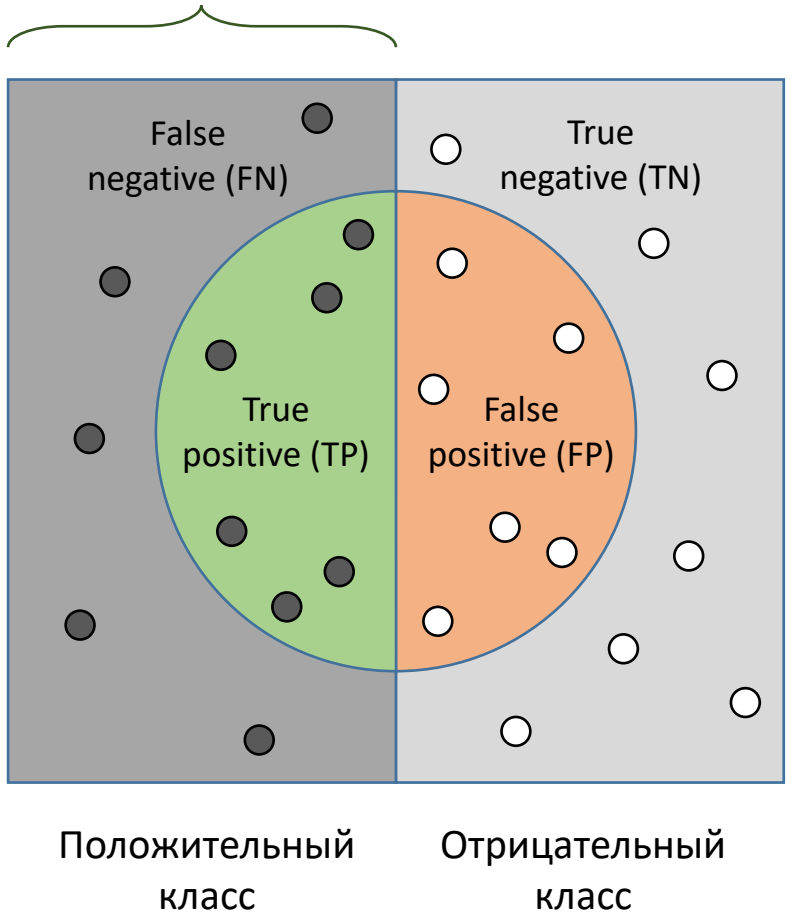
Матрица ошибок

		Истинная метка	
		Положительный класс	Отрицательный класс
Предсказанная метка	Положительный класс	TP	FP
	Отрицательный класс	FN	TN

Верность: $$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Точность: $$Precision = \frac{TP}{TP + FP}$$

Полнота: $$Recall = \frac{TP}{TP + FN}$$



Пример оценки для нескольких классов

accuracy: 96.00%

	true Iris-setosa	true Iris-versicolor	true Iris-virginica	class precision
pred. Iris-setosa	25	0	0	100.00%
pred. Iris-versicolor	0	23	1	95.83%
pred. Iris-virginica	0	2	24	92.31%
class recall	100.00%	92.00%	96.00%	

Матрица несоответствий для задачи с ирисами

Если классов больше чем два для получения точности и полноты применяется методика OvR (One versus Rest).

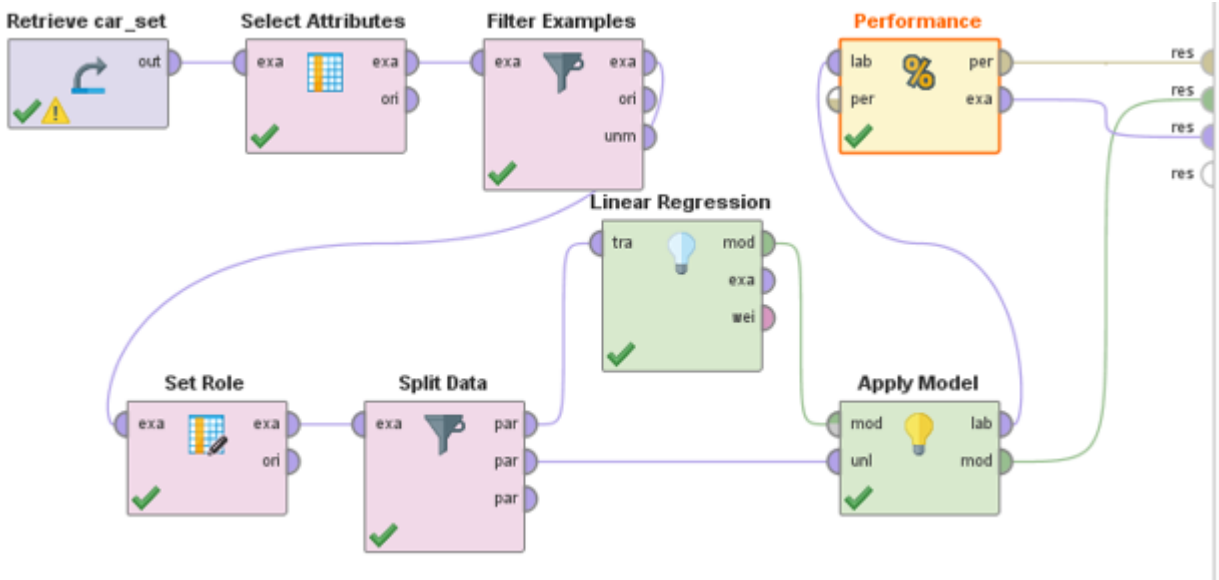
Для случаев трех классов: 1-й класс(+) против 2-й и 3-й классы(-),
2-й(+) против 1-й и 3-й классы(-), 3-й класс(+) против 1-й и 2-й классы(-)

Регрессия

Цель обучения: получить выражение зависимости типа Y=f(X), где Y – целевая переменная, а X – входные признаки.

Пример из https://archive.ics.uci.edu/ml/datasets/Auto+MPG
Выборка моделей автомобилей. Задача построить модель позволяющую оценить показатель Mpg (сколько миль проезжает автомобиль на галлоне топлива), т.е. 1/расход топлива

Набор входных параметров:
1. cylinders: Кол-во цилиндров двигателя
2. displacement: Объём двигателя
3. horsepower: Мощность двигателя
4. weight: Масса автомобиля
5. acceleration: Ускорение
6. model year: Год выпуска
7. car name: Наименование модели



Построение модели в среде Rapid Miner Studio

Оценка качества моделей регрессии

1.Средняя квадр. ошибка

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

2.Средняя абс. ошибка

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \tilde{y}_i|$$

3.Коэффициент детерминации

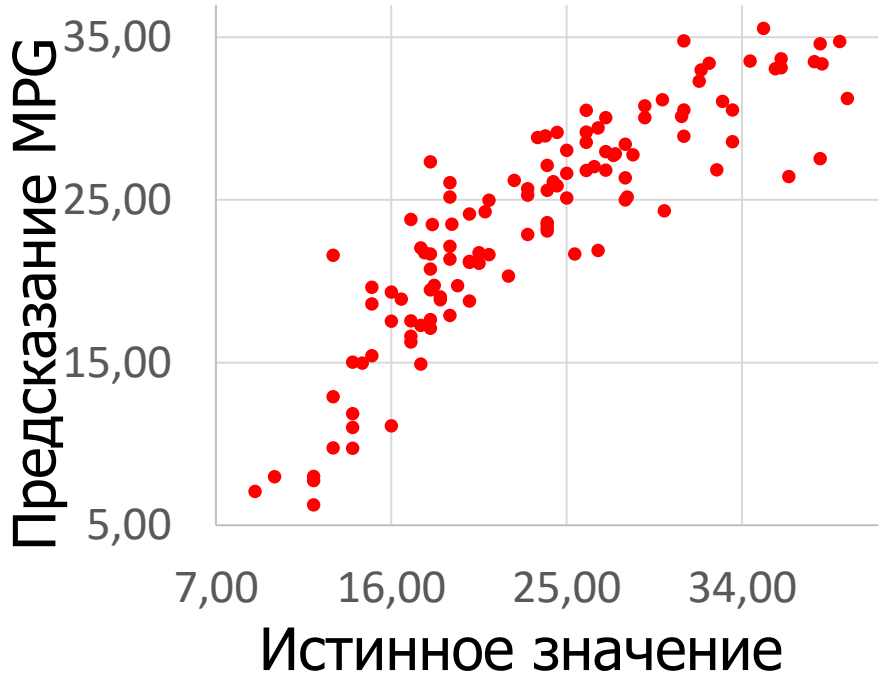
$$R2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

y_i - Истинное значение

$\bar{y}$ - Среднее значение

$\tilde{y}_i$ - Предсказанное значение

Row No.	MPG	prediction(MPG)
1	15	15.419
2	14	15.038
3	24	23.522
4	22	20.327
5	18	20.750
6	24	23.098
7	21	21.635
8	10	7.990
9	9	7.077
10	28	25.012
11	17	17.565
12	14	9.749
13	14	11.874
14	12	6.254
15	19	17.899
16	23	25.303



MSE = 3.48

R2 = 0.881

Обучение без учителя

Меток класса нет. Метод используется для изучения данных.

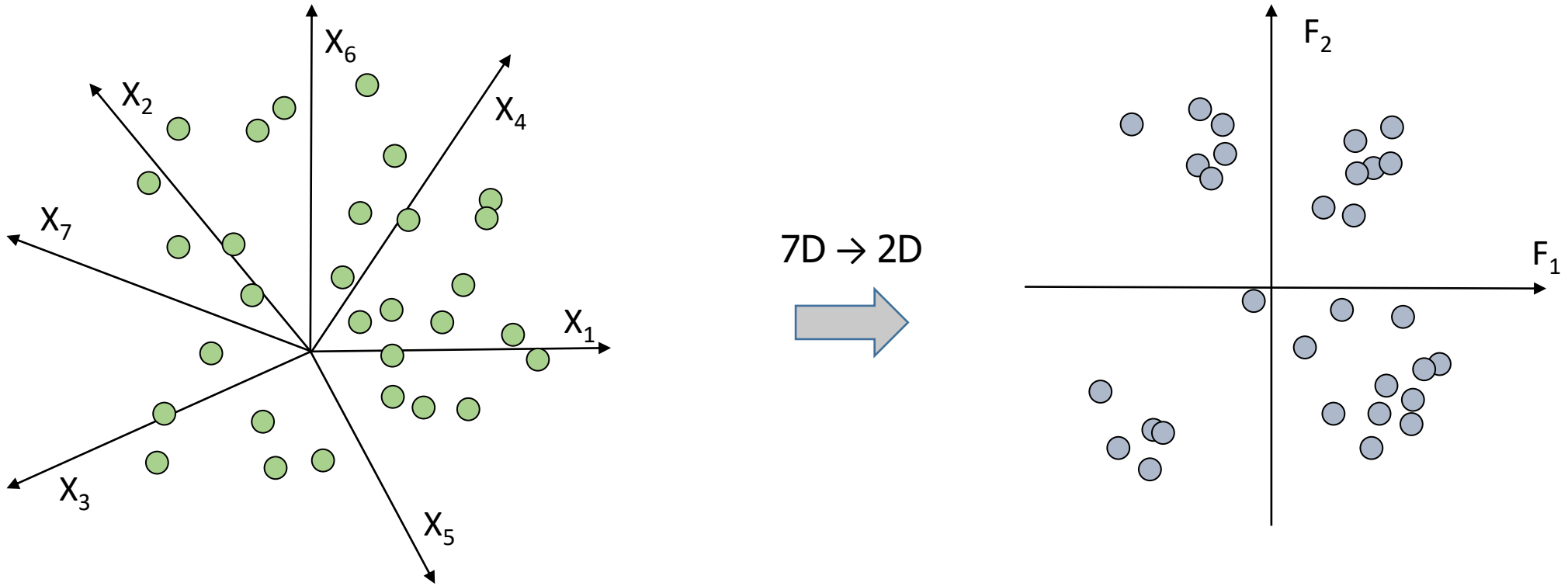
Применяется для задач, в которых известны описания множества объектов (обучающей выборки), и требуется обнаружить внутренние взаимосвязи, зависимости, закономерности, существующие между объектами.

- ✓ Кластеризация
- ✓ Снижение размерности
- ✓ Поиск аномалий

Снижение размерности признакового пространства

Преобразование данных из многомерного пространства в пространство с низкой размерностью, которое позволяет сохранить значимые свойства данных.

Уменьшение количества атрибутов при сохранении как можно большей вариации в оригинальном наборе.

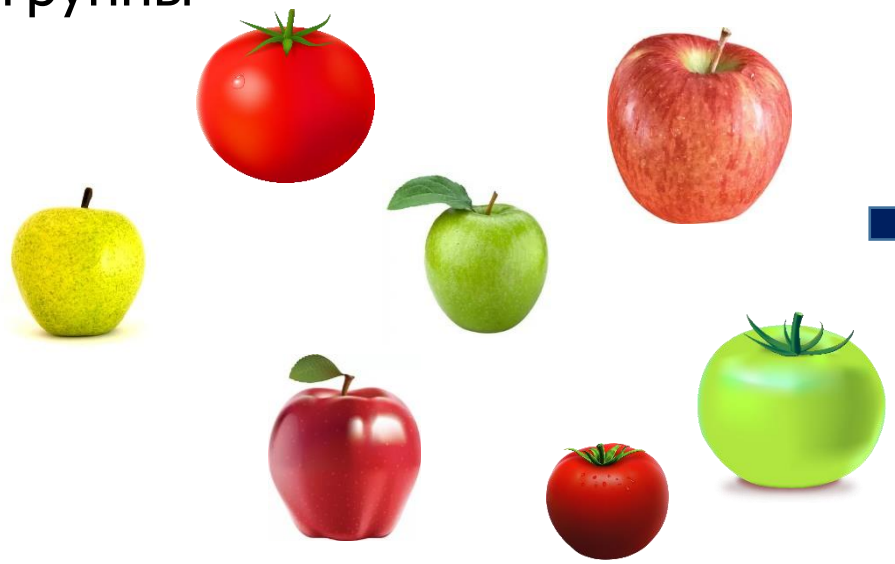


Кластеризация

Меток класса нет. Метод используется для изучения данных.

Особенность: Субъективность кластеризации.

Задача: Разложить объекты на две группы



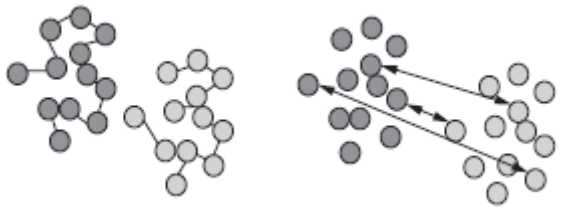
Оценка качества моделей кластеризации

Индексы качества кластеризации.
Оценка производится методом сравнения нескольких структур

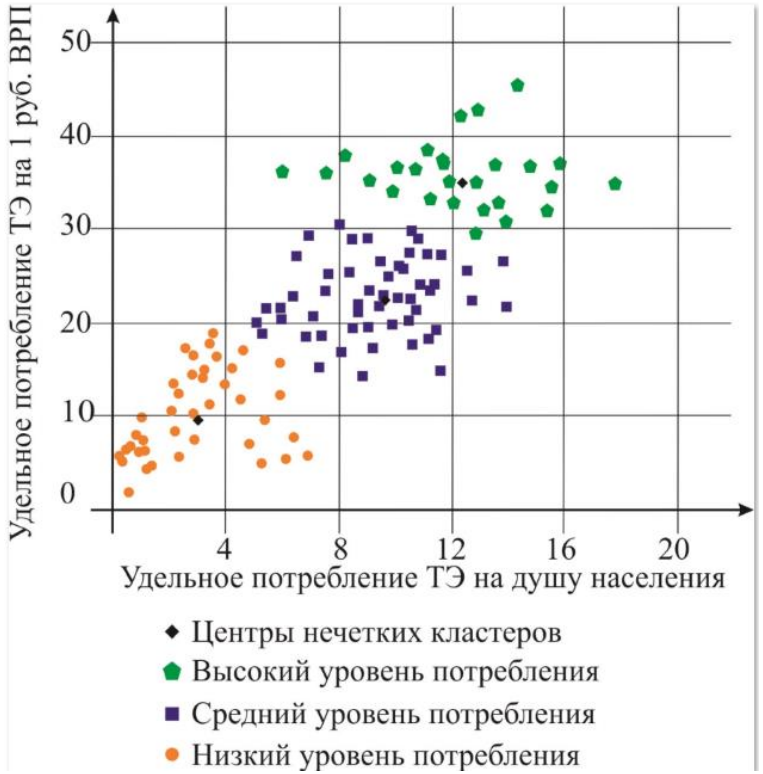
- Нескольких запусков одного и того же алгоритма
- Запуск алгоритма с разными параметрами
- Запуск разных алгоритмов

Критерии оценки качества:

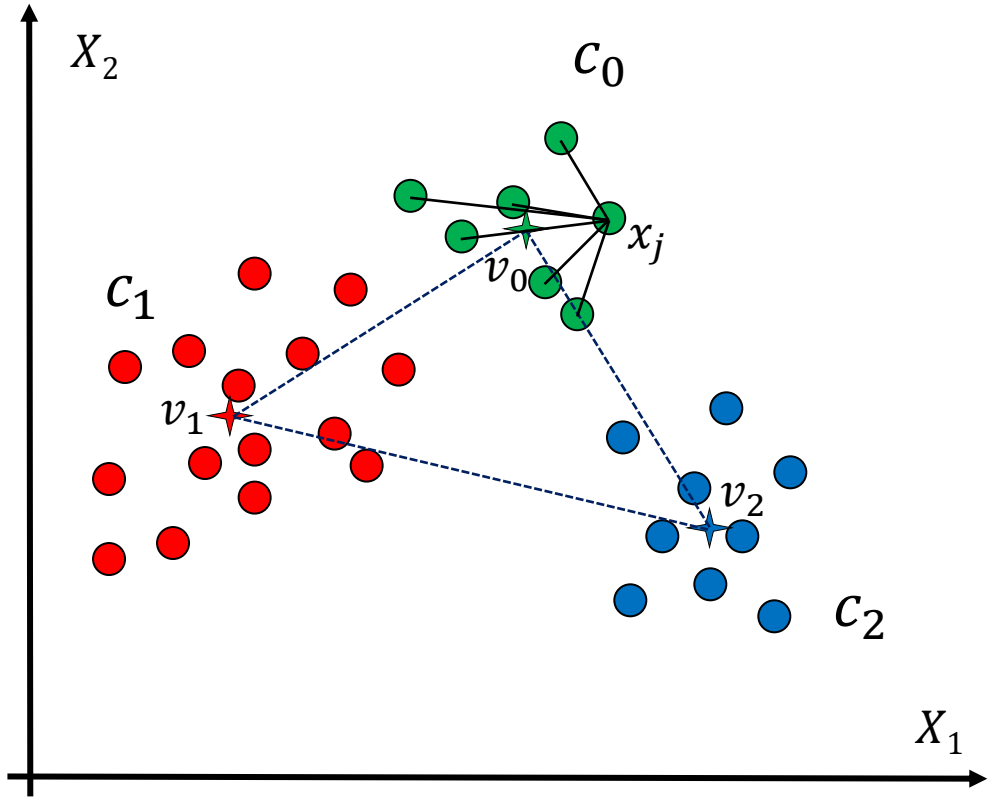
Компактность - элементы из одного кластера должны быть как можно ближе друг к другу.
Отделимость – элементы из разных кластеров должны быть как можно дальше друг от друга.

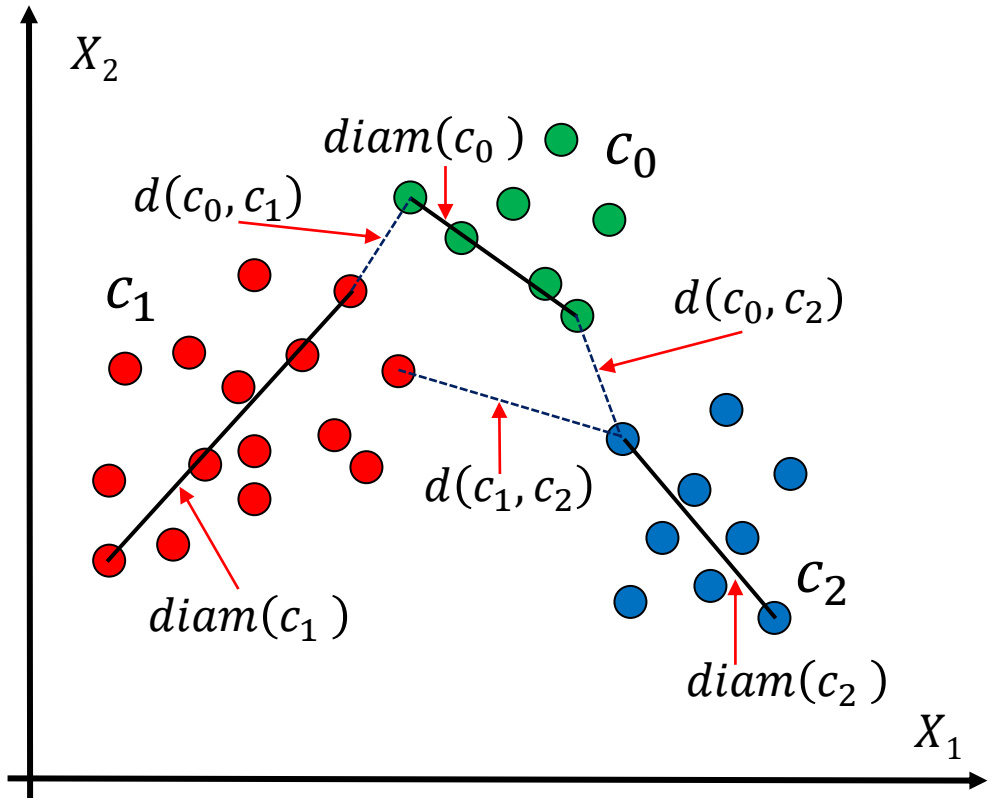


Пример кластеризации

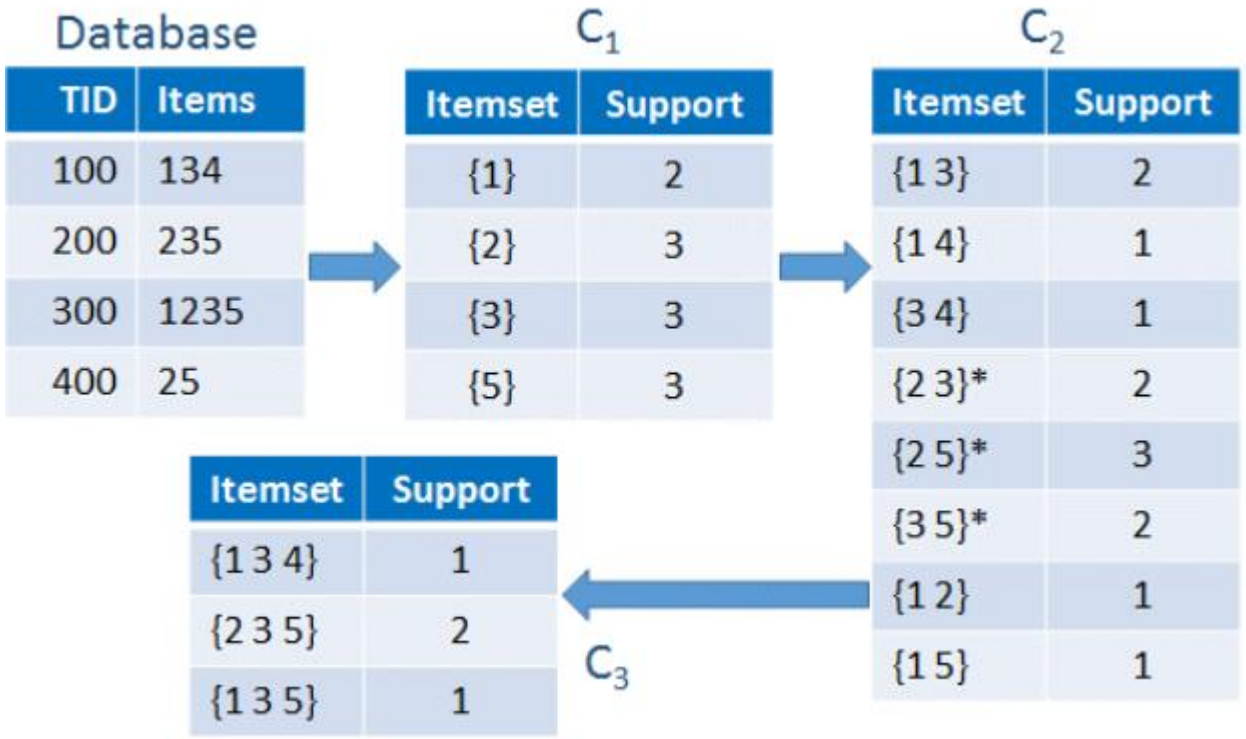


Внутри- и меж-кластерное расстояния

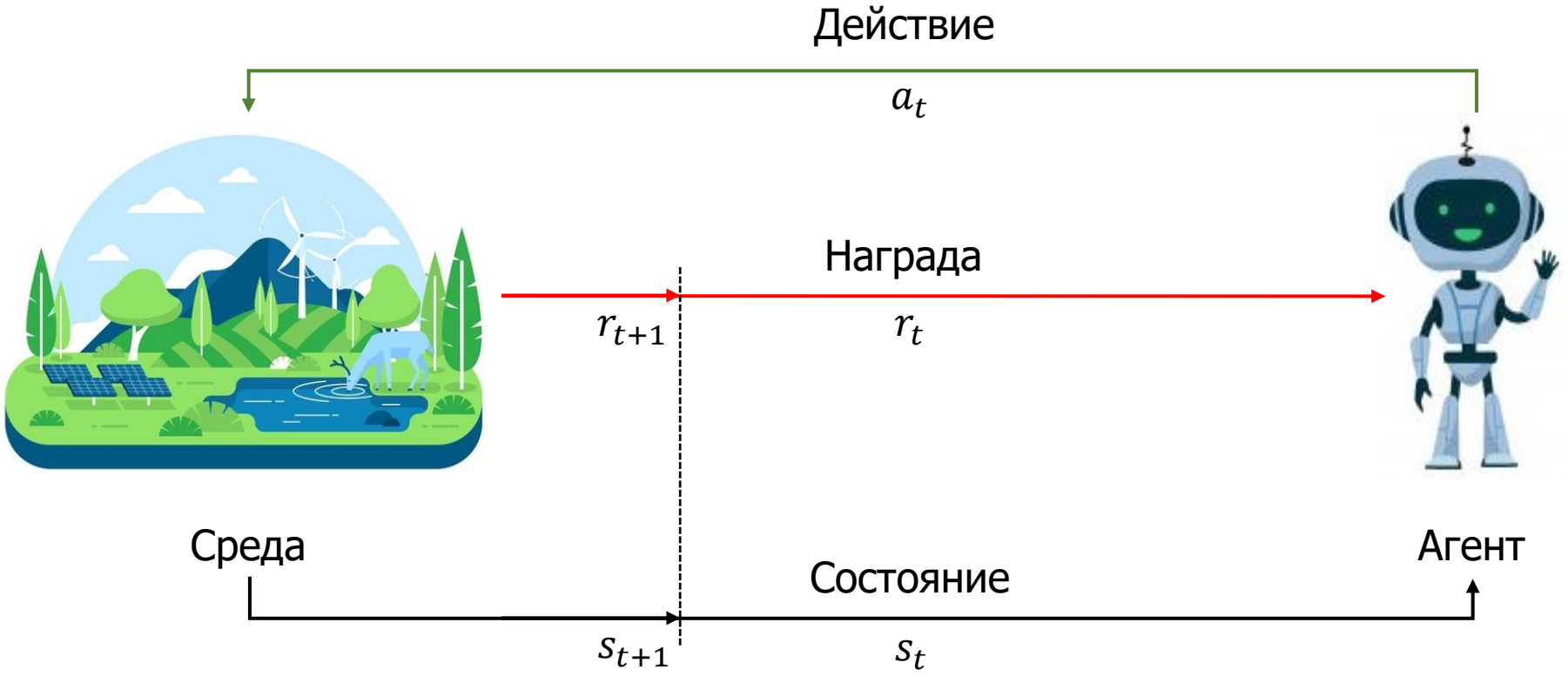




Ассоциации


Database
TID
Items
100
134
200
235
300
1235
400
25
C1
Itemset
Support
{1}
2
{2}
3
{3}
3
{5}
3
C2
Itemset
Support
{1 3}
2
{1 4}
1
{3 4}
1
{2 3}*
2
{2 5}*
3
{3 5}*
2
{1 2}
1
{1 5}
1
C3
Itemset
Support
{1 3 4}
1
{2 3 5}
2
{1 3 5}
1

Обучение с подкреплением



Использованы изображения из freepik