

Лекция №1

Введение в машинное обучение

Спасёнов Алексей

Введение в машинное обучение



Преподаватели:



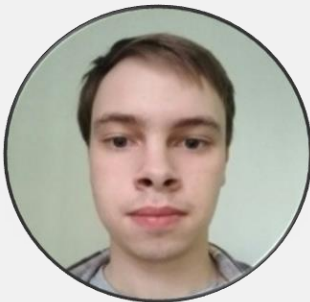
Спасёнов Алексей

Программист-исследователь,
Группа предиктивной аналитики



Некрасов Евгений

Старший программист-исследователь,
Группа предиктивной аналитики



Логачёв Павел

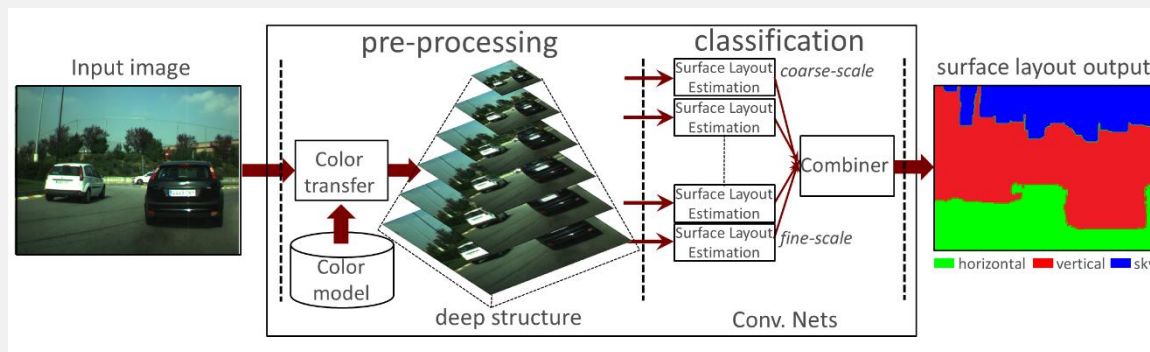
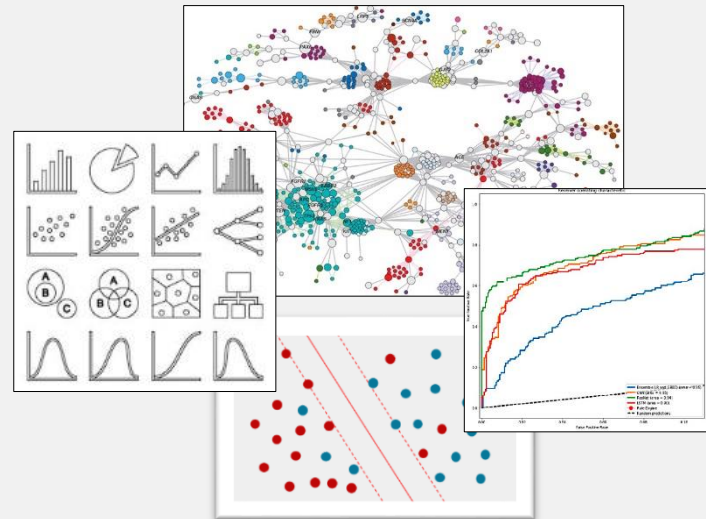
Программист-исследователь,
Группа персонализации

Введение в машинное обучение



Что будет в курсе:

Name	HR Information			Contact	
	Position	Salary	Office	Office	Extn.
Airi Satou	Accountant	\$162,700	Tokyo		5407
Angelica Ramos	Chief Executive Officer (CEO)	\$1,200,000	London		5797
Ashton Cox	Junior Technical Author	\$86,000	San Francisco		1562
Bradley Greer	Software Engineer	\$132,000	London		2558
Brenden Wagner	Software Engineer	\$206,850	San Francisco		1314
Brielle Williamson	Integration Specialist	\$372,000	New York		4804
Bruno Nash	Software Engineer	\$163,500	London		6222
Caesar Vance	Pre-Sales Support	\$106,450	New York		8330
Cara Stevens	Sales Assistant	\$145,600	New York		3990
Cedric Kelly	Senior Javascript Developer	\$433,060	Edinburgh		6224
Name	Position	Salary	Office	Office	Extn.





Содержание курса

1. Введение в анализ данных и машинное обучение
2. Статистические методы анализа данных
3. Обучение без учителя
4. Задачи классификации и регрессии
5. Ансамбли моделей
6. Преобразование и получение признаков
7. Работа с текстовыми данными
8. Нейронные сети. Свёрточные нейронные сети.
Рекуррентные нейронные сети
9. Обучение с подкреплением

Введение в машинное обучение. Лекция 1



Выполнение заданий:

1) Отправляем решения на
a.spasenov@corp.mail.ru
e.nekrasov@corp.mail.ru

2) В теме письма добавляем:
[TSMLintro]

Например:

[TSMLintro] Contest 1

Введение в машинное обучение. Лекция 1



Содержание лекции

1. Основный понятия
2. Основные типы задач
3. Примеры прикладных задач
4. Знакомство с библиотеками:
 1. Numpy
 2. Matplotlib
 3. Pandas
5. Практика: анализ банковских данных

Рекомендуемая литература



- Christopher M. Bishop. Pattern recognition and Machine Learning
- Kevin P. Murphy. Machine Learning. A Probabilistic Perspective
- Ian Goodfellow, Yoshua Bengio, Aaron Courville. Deep Learning

Технострим Mail.ru Group:

1. Введение в анализ данных
2. Data Mining
3. Методы обработки больших объёмов данных



Машинное обучение (Machine Learning)

Обширный подраздел прикладной математики, находящийся на стыке математической статистики, оптимизации, искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться по эмпирическим (прецедентным) данным.

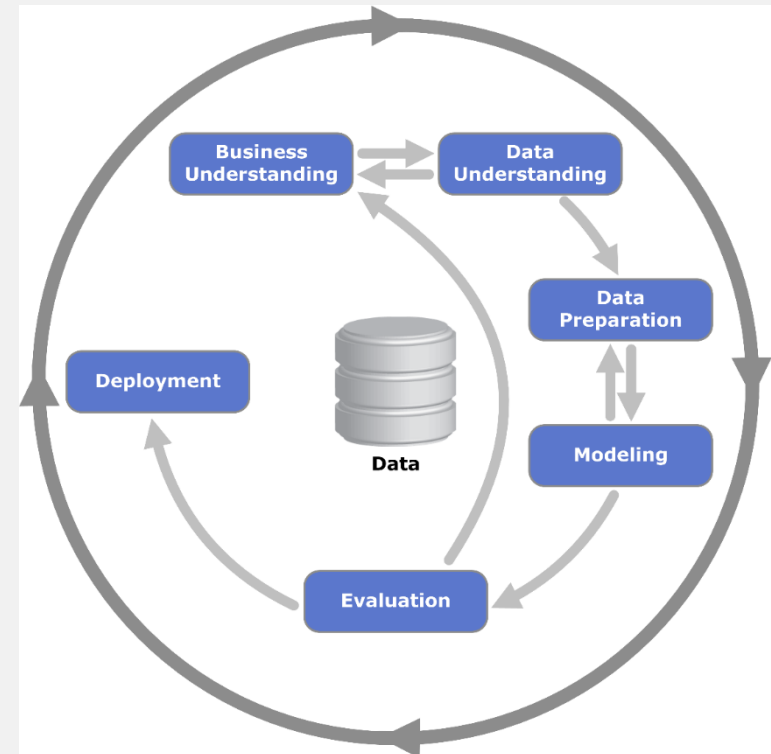


Анализ данных (Data Mining)

Процесс извлечения знаний из различных источников данных, таких как базы данных, текст, картинки, видео и т.д. Полученные знания должны быть достоверными, полезными и интерпретируемыми.



Cross Industry Standard Process for Data Mining

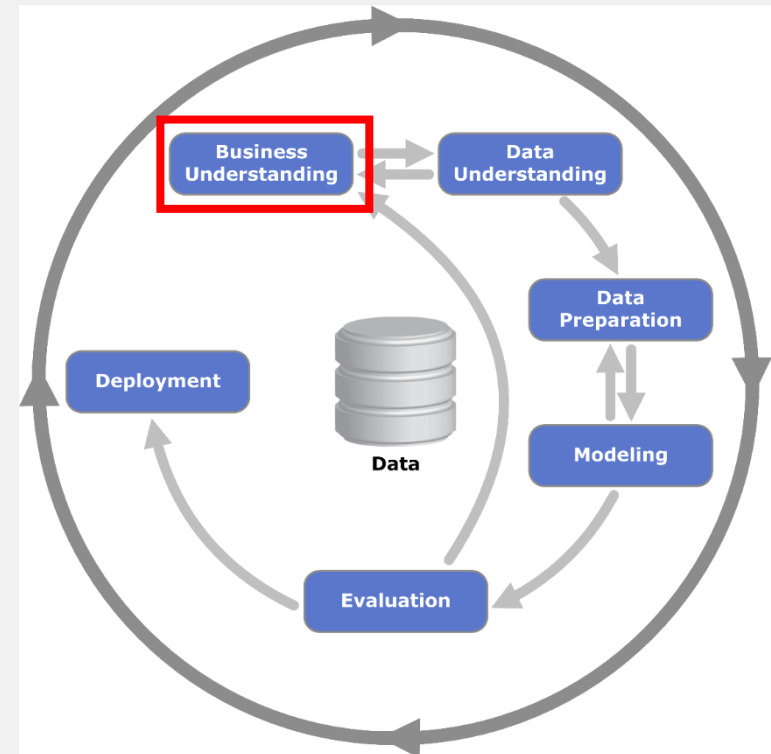




Постановка задачи

- 1) Предсказание оттока клиентов с сайта
- 2) Распознавание марки и модели автомобилей по изображениям
- 3) Информационный поиск, анализ текстов
- 4) Кредитный скоринг
- 5) Социологические исследования
- 6) Медицинская диагностика

...



Введение в машинное обучение



Признаки (Features)

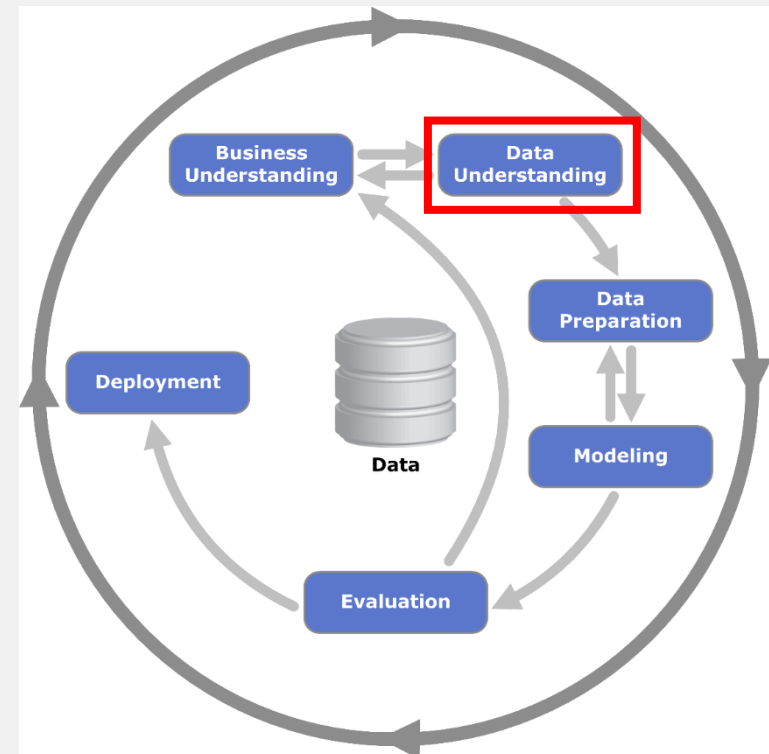
D – множество объектов (Data set)

$d \in D$ – обучающий объект

$\phi_i : D \rightarrow F_j$ – признак

Виды признаков:

1) Бинарные	Binary	$F_j = \{true, false\}$
2) Номинальные	Categorical	F_j – конечно
3) Порядковые	Ordinal	F_j – конечно упорядочено
4) Количественные	Numerical	$F_j = \mathbb{R}$





Пример:

Задача: Необходимо спрогнозировать стоимость дома

Признаки, характеризующие стоимость жилья:

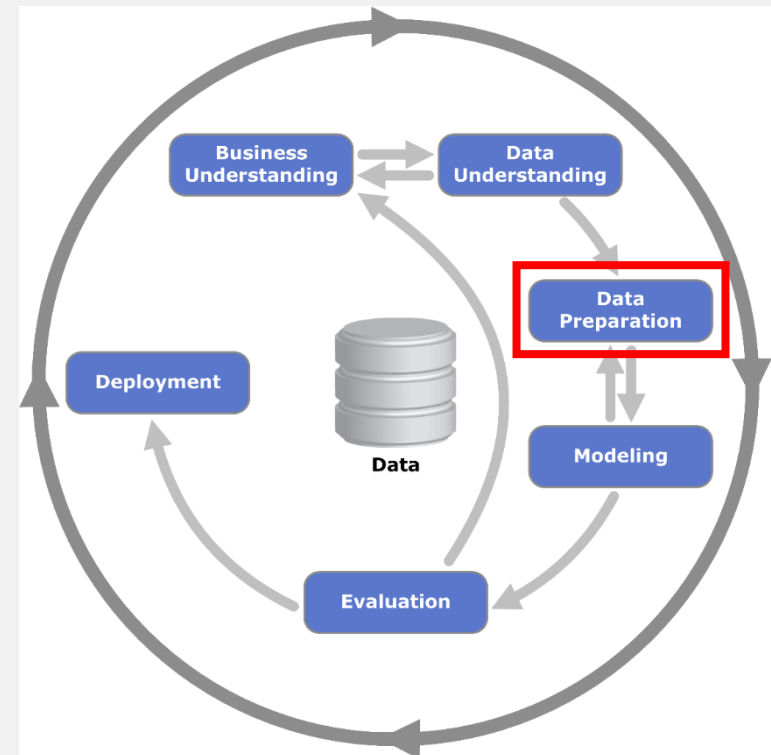
Бинарные	Номинальные	Порядковые	Количественные
Наличие отсутствие газа (электричества) Наличие отсутствие подвального помещения	Регион расположения	Число владельцев Число комнат Число этажей	Удалённость от общественного транспорта Удалённость от водоёма



Подготовка данных

(Data Preparation)

- 1) Удаление шума
- 2) Заполнение отсутствующих значений
- 3) Трансформация значений
- 4) Выбор факторов
- 5) Использование априорных знаний



```
graph TD; BU[Business Understanding] <--> DU[Data Understanding]; DU --> DP[Data Preparation]; DP <--> M[Modeling]; M --> E[Evaluation]; E --> D[Deployment]; D --> BU; Data[(Data)]
```

The diagram illustrates the Machine Learning lifecycle as a continuous loop. The stages are: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The 'Modeling' stage is highlighted with a red border. A central database icon is labeled 'Data'.



Основные типы задач

1) Обучение с учителем (supervised learning)

Каждый прецедент представляет собой пару «объект, ответ». Требуется найти функциональную зависимость ответов от описаний объектов и построить алгоритм, принимающий на входе описание объекта и выдающий на выходе ответ.

2) Обучение без учителя (unsupervised learning)

Ответы не задаются, и требуется искать зависимости между объектами

3) Частичное обучение (semi-supervised learning)

Каждый прецедент представляет собой пару «объект, ответ», но ответы известны только на части прецедентов.

4) Обучение с подкреплением (reinforcement learning)

Роль объектов играют пары «ситуация, принятое решение», ответами являются значения функционала качества, характеризующего правильность принятых решений (реакцию среды).

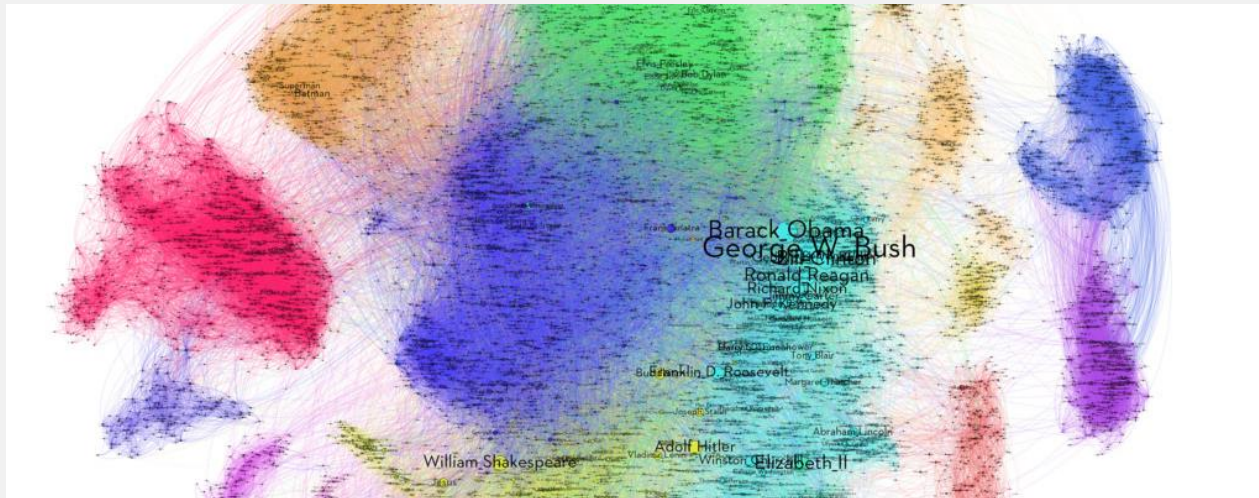
...

Введение в машинное обучение



Пример обучение без учителя

Поиск документов (статей, сайтов т.д.) имеющих похожую тематику



Введение в машинное обучение



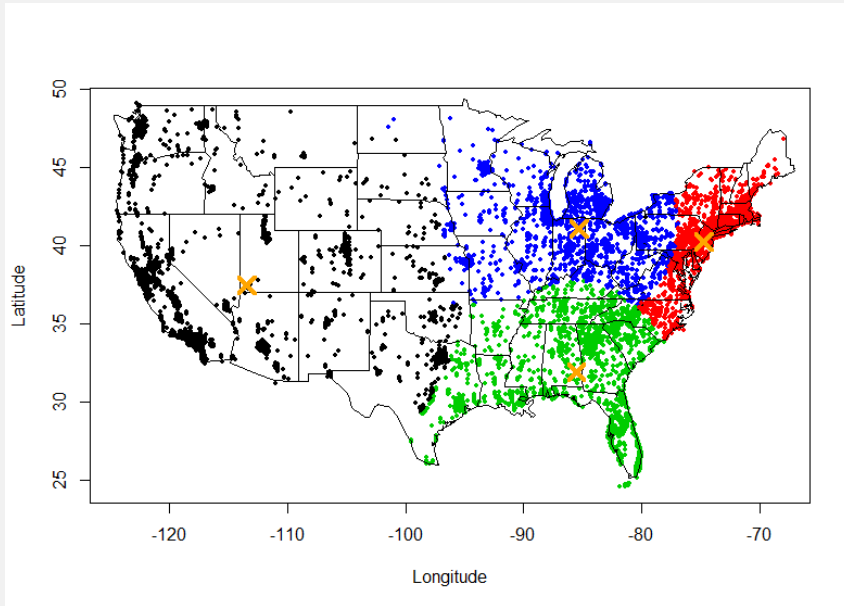
Пример обучение без учителя

Поиск музыки одинакового жанра

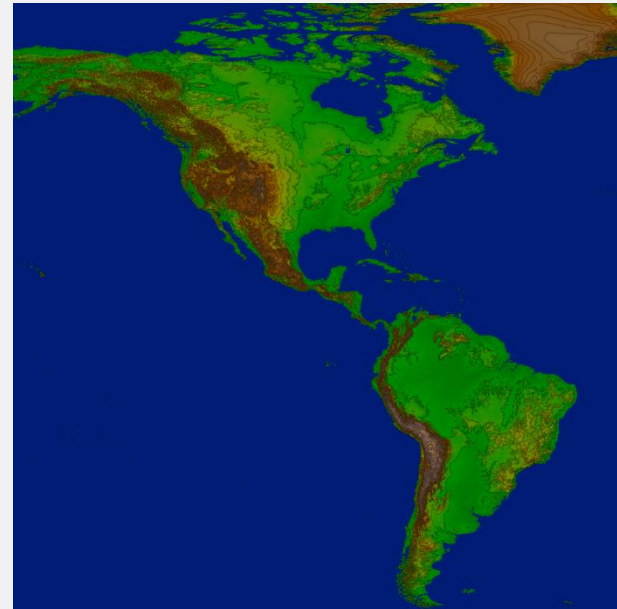




Пример обучение без учителя



Уровень заработка



Регионы сейсмической активности



Обучение с учителем (обучения по прецедентам)

Модель

Семейство параметрических функций вида

$$H = \{h(x, \Theta): \mathcal{X} \times \Theta \rightarrow Y\}$$

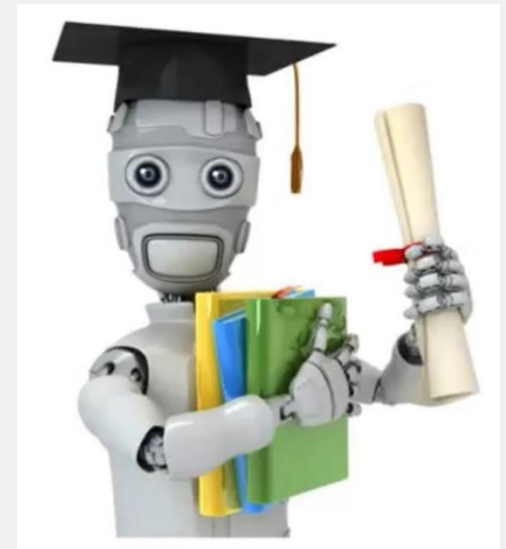
Алгоритм обучения

Выбор наилучших параметров Θ

$$A(X, Y): (X \times Y)^N \rightarrow \Theta$$

В итоге:

$$h^*(x) = h(x, \Theta^*)$$





Обучение с учителем (обучения по прецедентам)

Задачи классификации (classification)

- $F_j = \{true, false\}$ – классификация на 2 класса
- $F_j = \{1, \dots, M\}$ – классификация на M непересекающихся классов
- $F_j = \{0,1\}^M$ – классификация на M классов, которые могут пересекаться

Задача восстановления регрессии (regression)

- $F_j = \mathbb{R}$ или $F_j = \mathbb{R}^M$ (ответом является действительное число или числовой вектор)

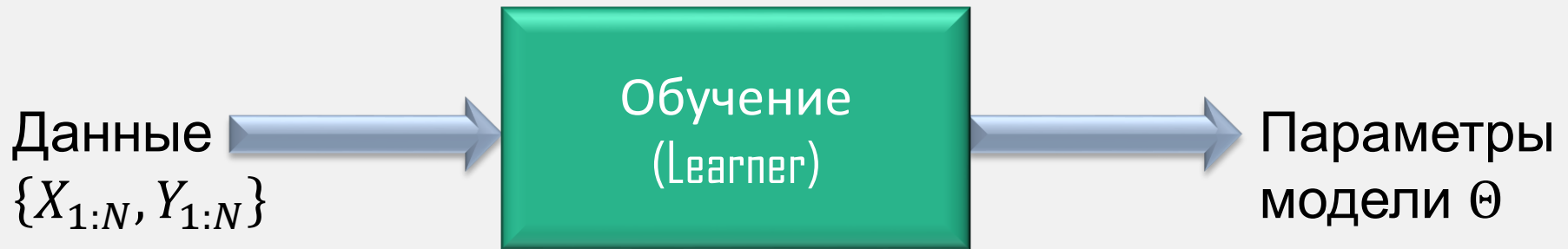
Задача ранжирования (learning to rank)

- F_j – конечно упорядочено (ответы надо получить сразу на множестве объектов, после чего отсортировать их по значениям ответов)



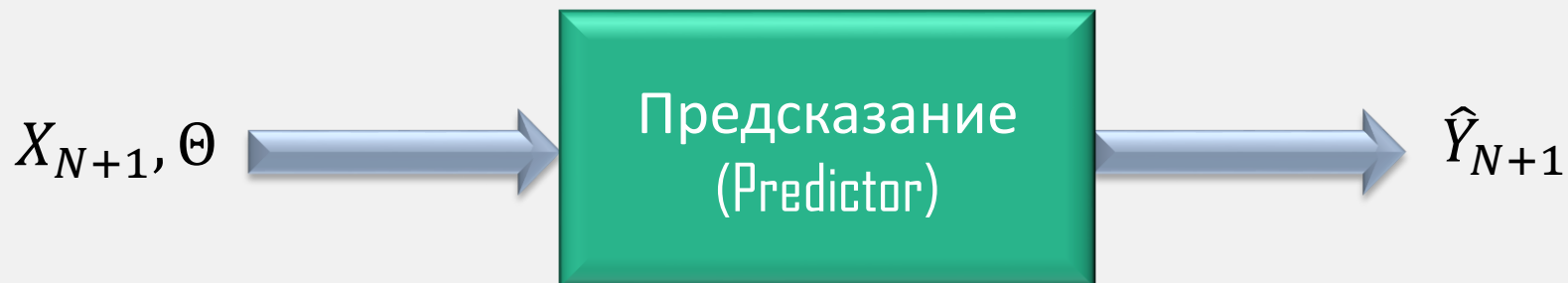
Обучение с учителем (обучения по прецедентам)

Этап обучения (train)



Необходимо учитывать представительность выборки

Этап применения (test)

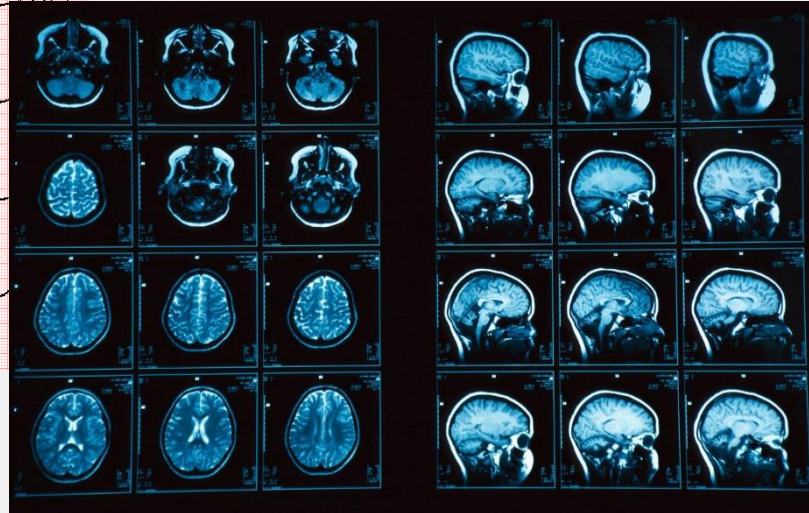
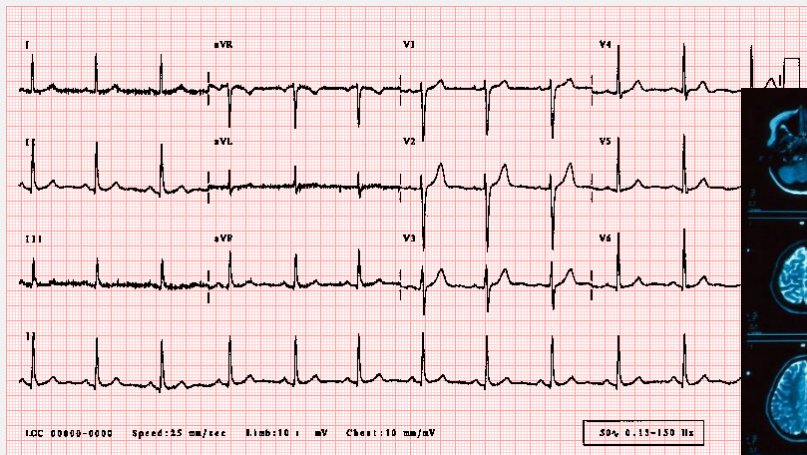




Обучение с учителем (обучения по прецедентам)

Пример задачи классификации

Пациент здоров или болен? (бинарная классификация)





Обучение с учителем (обучения по прецедентам)

Пример задачи классификации

Какая марка автомобиля? (классификация на M непересекающихся классов)





Обучение с учителем (обучения по прецедентам)

Пример задачи классификации

Что представлено на изображении? (классификация на M классов, которые могут пересекаться)





Обучение с учителем (обучения по прецедентам)

Пример задачи классификации

Психитипирование личности (BIG5, MBTI) (классификация на M классов, которые могут пересекаться)





Обучение с учителем (обучения по прецедентам)

Пример задачи восстановления регрессии

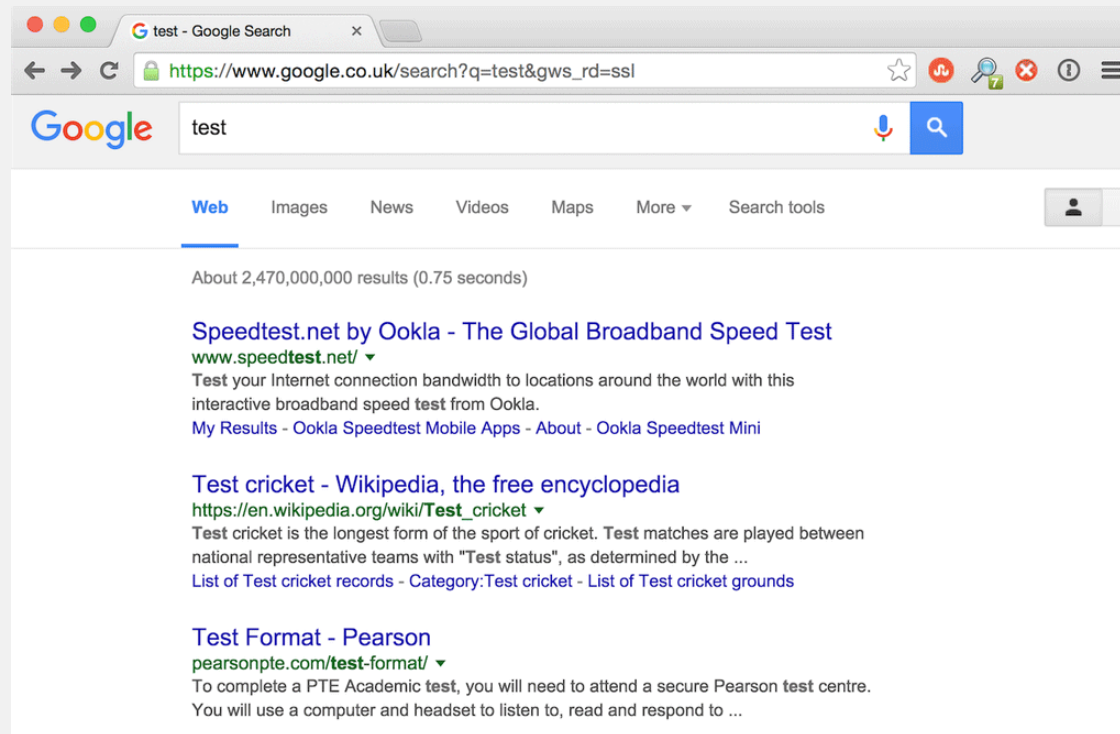
Предстаказание стоимости валют





Обучение с учителем (обучения по прецедентам)

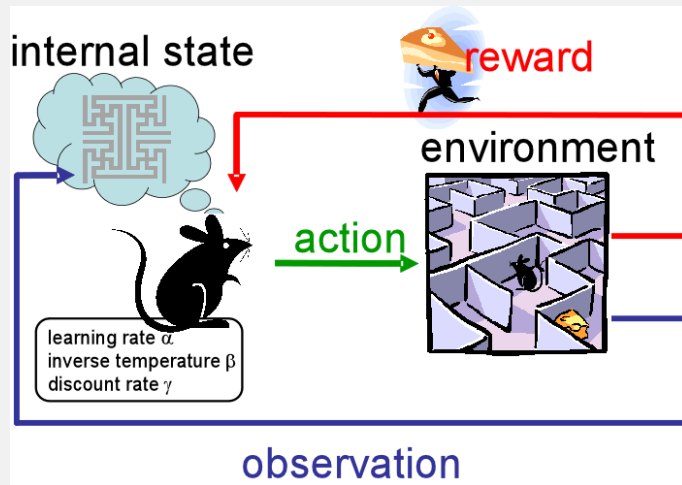
Пример: задачи ранжирования





Обучение с подкреплением (reinforcement learning)

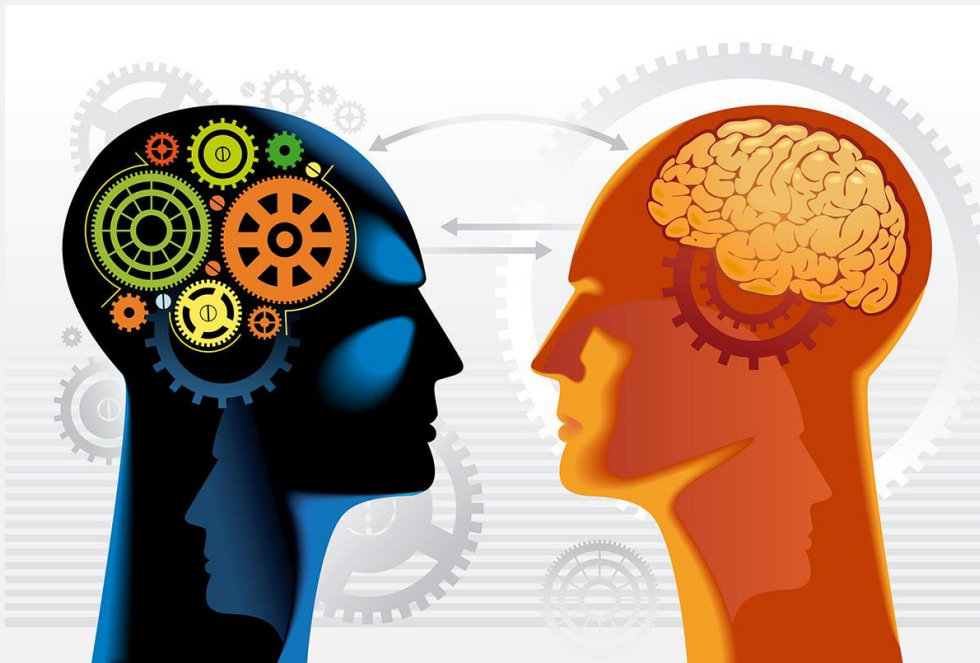
Пример: игры atari





Обучение с подкреплением (reinforcement learning)

Пример: создание чатбота





Инструменты



Введение в машинное обучение



Инструменты





**Спасибо за
внимание!**

Спасёнов Алексей

a.spasenov@corp.mail.ru