



# Введение в машинальное обучение

**Лекция 1**

*Введение в анализ данных и машинное обучение*

**Журавлёв Вадим**

# Преподаватели



**Вадим Журавлёв**

Программист-исследователь  
Группа data-аналитики  
Mail.ru Group



**Мамаев Александр**

Руководитель группы  
Группа data-аналитики  
Mail.ru Group



**Георгий Господинов**

Руководитель команды  
Группа контекста и потоковой  
обработки  
Mail.ru Group



**Андрей Шестаков**

Программист-исследователь  
Profi.ru



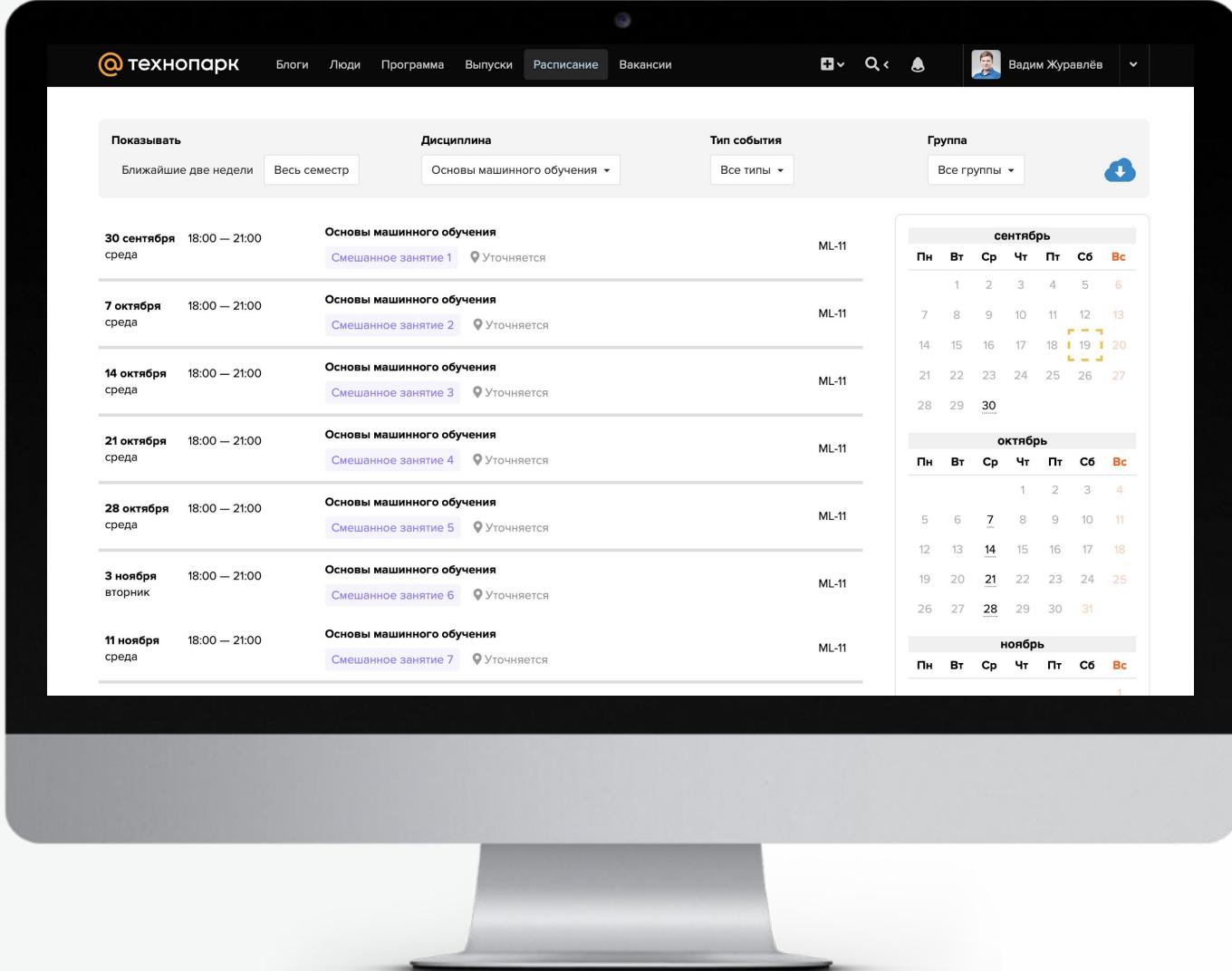
**Шематоров Константин**

Программист-  
исследователь  
Группа персонализации  
Mail.ru Group

---

## **Предполагаемые навыки:**

1. Основы линейной алгебры (базовые операции, свойства векторов и матриц, такие как скалярное произведение, норма вектора, обусловленность матрицы);
2. Основы математического анализа (производные и интегралы);
3. Теория вероятностей и математическая статистика (понятия вероятности, условной вероятности, функции распределения);
4. Python (хотя бы чуть-чуть).



Не забываем  
отмечаться на  
портале и  
оставлять  
отзывы!

---

# Содержание курса

1. Введение в анализ данных и машинное обучение
2. Задачи классификации и регрессии I
3. Задачи классификации и регрессии II
4. Оценка качества моделей и работа с признаками
5. Работа с текстовыми данными I
6. Обучение без учителя
7. Ансамбли моделей
8. Работа с текстовыми данными II
9. Рекомендательные системы
10. Работа с гео-данными
11. Анализ графов
12. Анализ сигналов
13. Экзамен

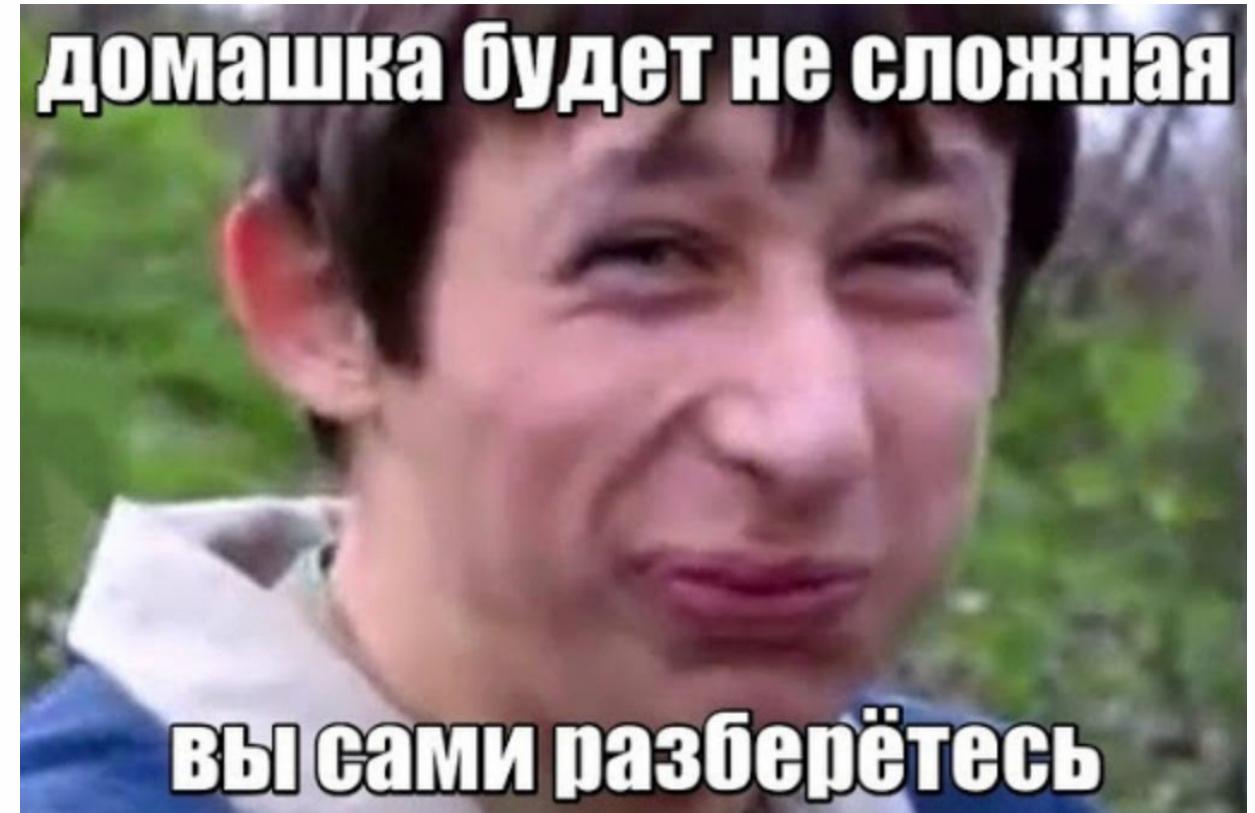
---

# Содержание курса

1. Введение в анализ данных и машинное обучение
2. Задачи классификации и регрессии I
3. Задачи классификации и регрессии II
4. Оценка качества моделей и работа с признаками
5. Работа с текстовыми данными I
6. Обучение без учителя
7. Ансамбли моделей
8. Работа с текстовыми данными II
9. Рекомендательные системы
10. Работа с гео-данными
11. Анализ графов
12. Анализ сигналов
13. Экзамен

## Выполнение заданий

- 1) Задания будут выкладываться на портал
- 2) Решения тоже нужно заливать на портал в виде ссылки на google colab



# **Домашние задания**

---

В курсе будет 11 домашних заданий + экзамен

Правила сдачи ДЗ:

- На каждое задание будет дедлайн
- После окончания дедлайна решения можно присыпать в течение недели, каждый день просрочки - минус балл
- Спустя неделю решения не принимаем
- Мы против плагиата - первый плагиат - всем замеченным 0 баллов за ДЗ, второй плагиат - отчисление с курса

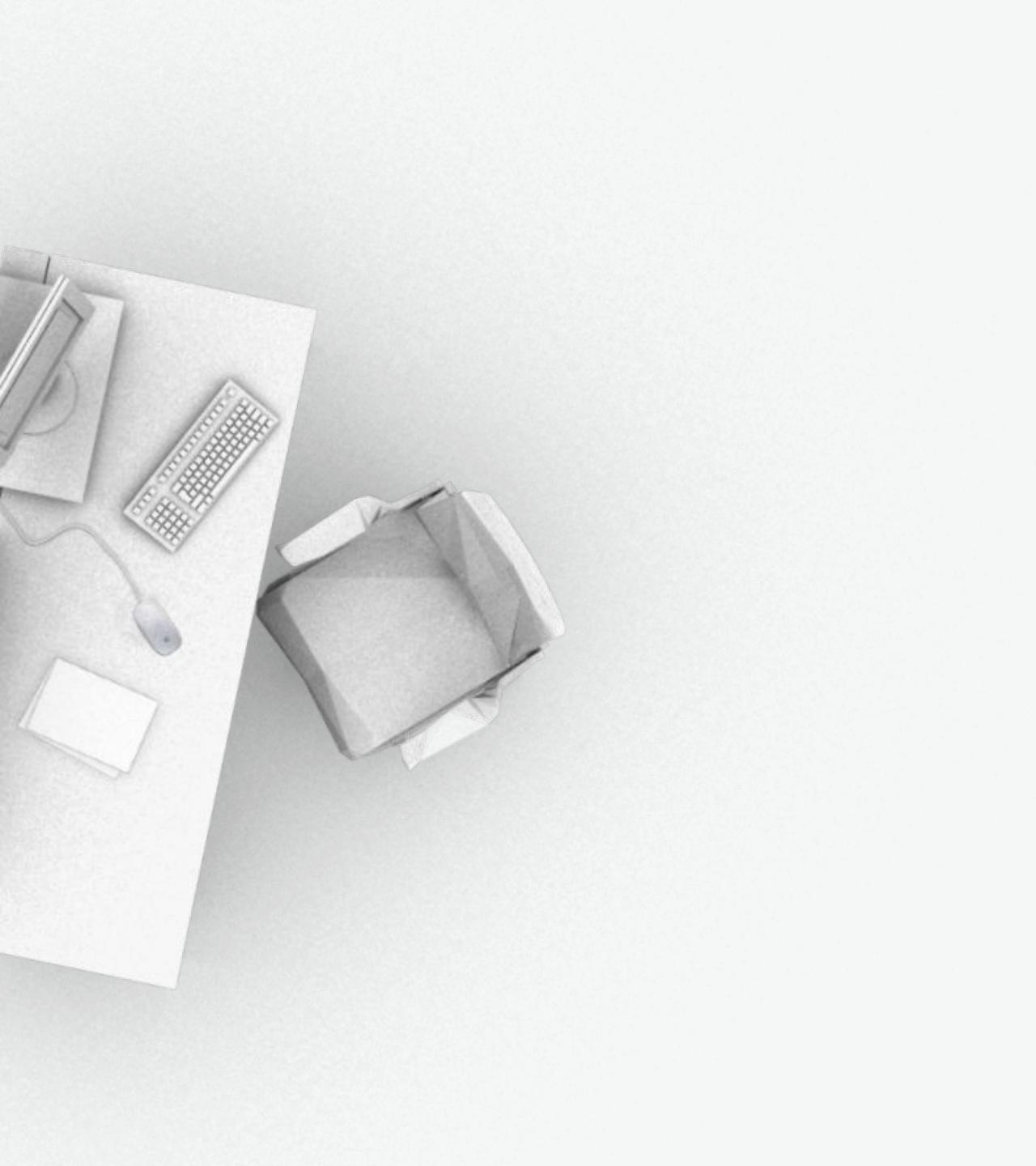
# **Экзамен**

---

Что входит в экзамен?

Будет необходимо приготовить  
презентацию-отчёт по всем ДЗ

На одно ДЗ – 1-2 слайда



# **Введение в машинное обучение**

Лекция 1

---

## **Содержание лекции**

1. Основные понятия
2. Основные типы задач
3. Примеры прикладных задач
4. Знакомство с библиотеками:
  1. Numpy
  2. Pandas
  3. Matplotlib

# 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

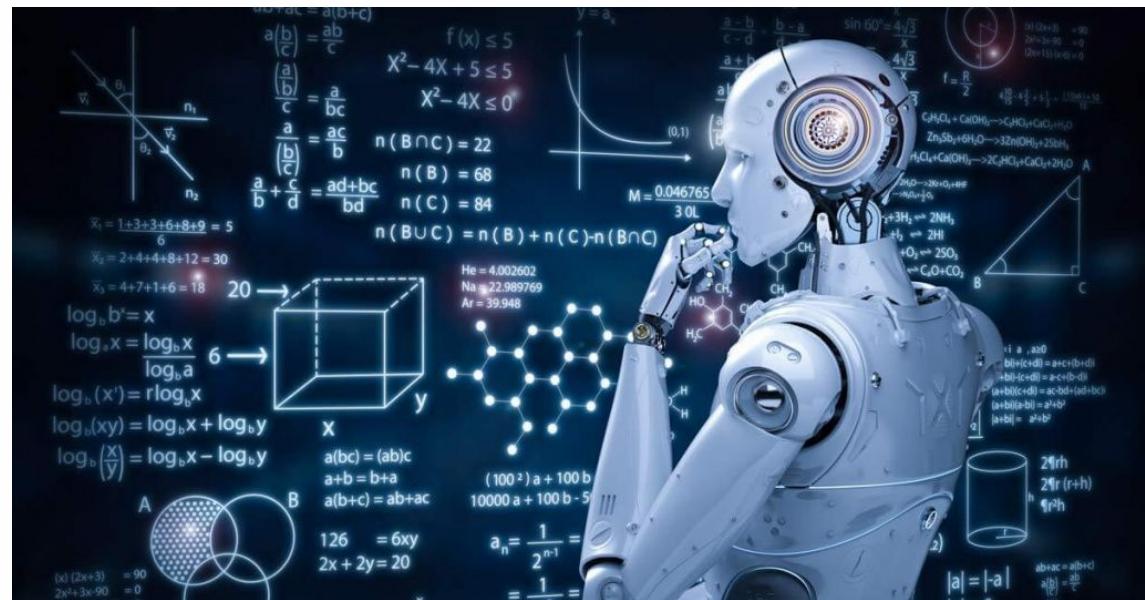
**Технострим Mail.Ru Group:**

<https://www.youtube.com/user/TPMGTU>

Технопарк, Техносфера, Технотрек, Техноатом, Технополис

# 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.

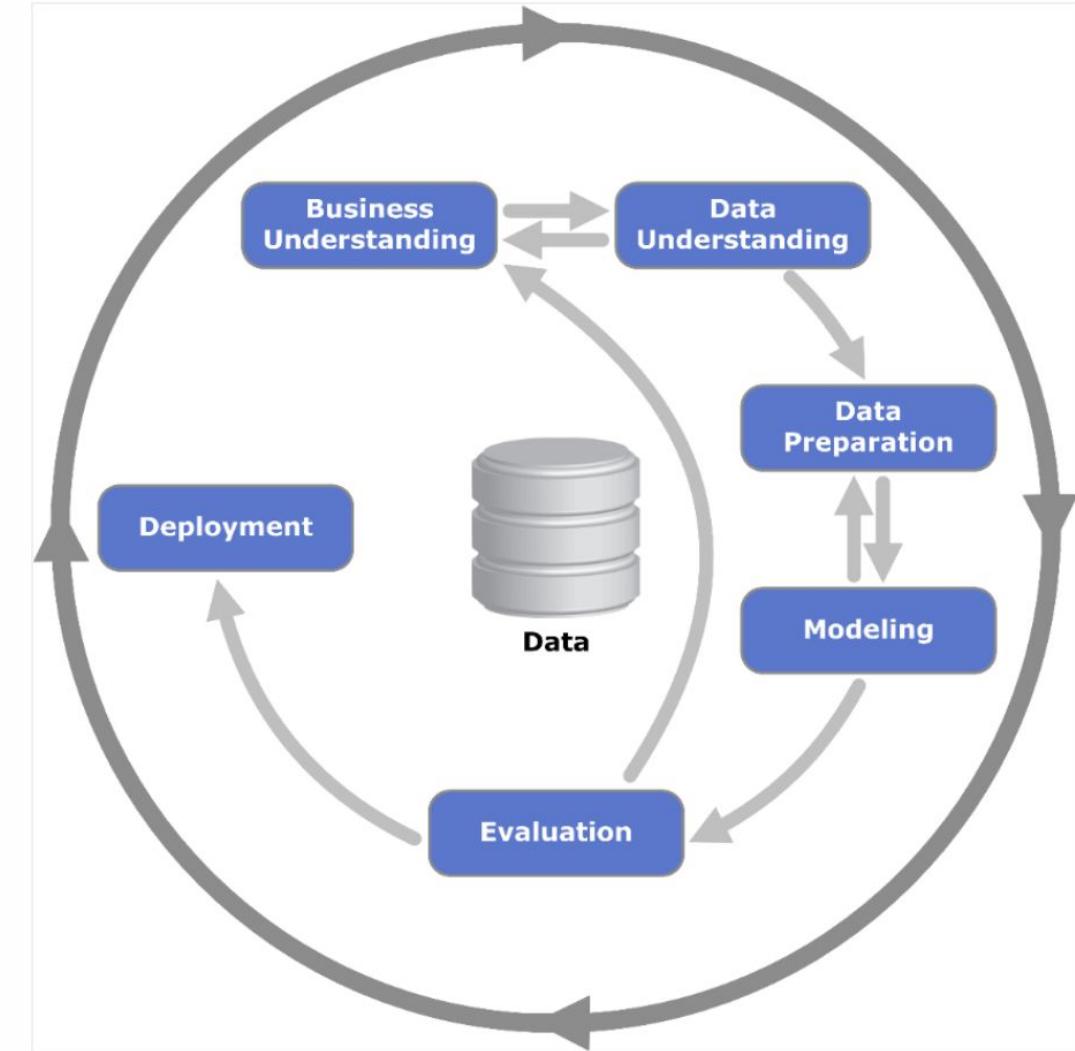


## Анализ данных (Data Mining)

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.



# CRISP-DM: Cross Industry Standard Process for Data Mining



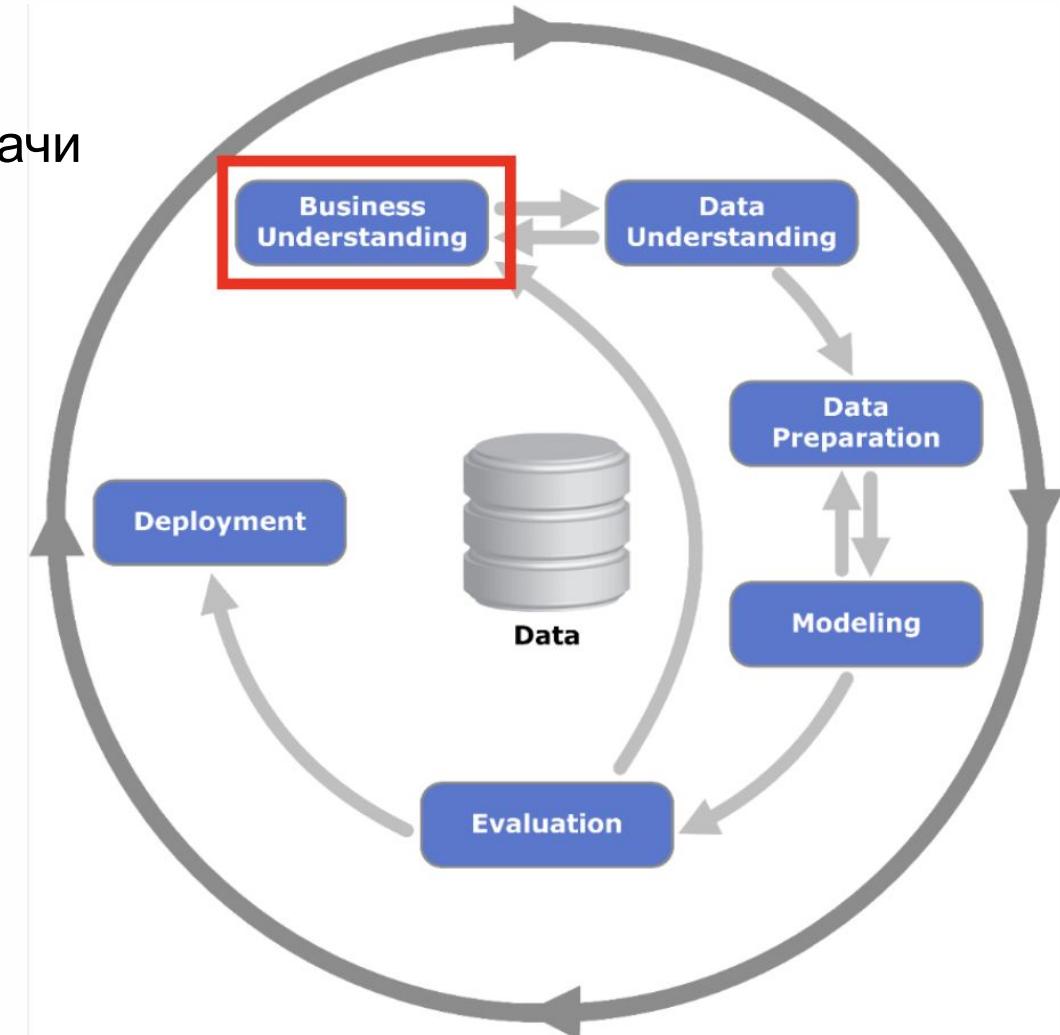
# Бизнес-анализ:

- определение бизнес целей - постановка задачи
- оценка текущей ситуации
- определение целей аналитики
- подготовка плана проекта

## Постановка задачи

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

...



# Анализ данных

- Сбор данных
- Описание данных
- Изучение данных
- Проверка качества данных



## Признаки (Features)

$D$  – множество объектов (Data set)

$d \in D$  – обучающий объект

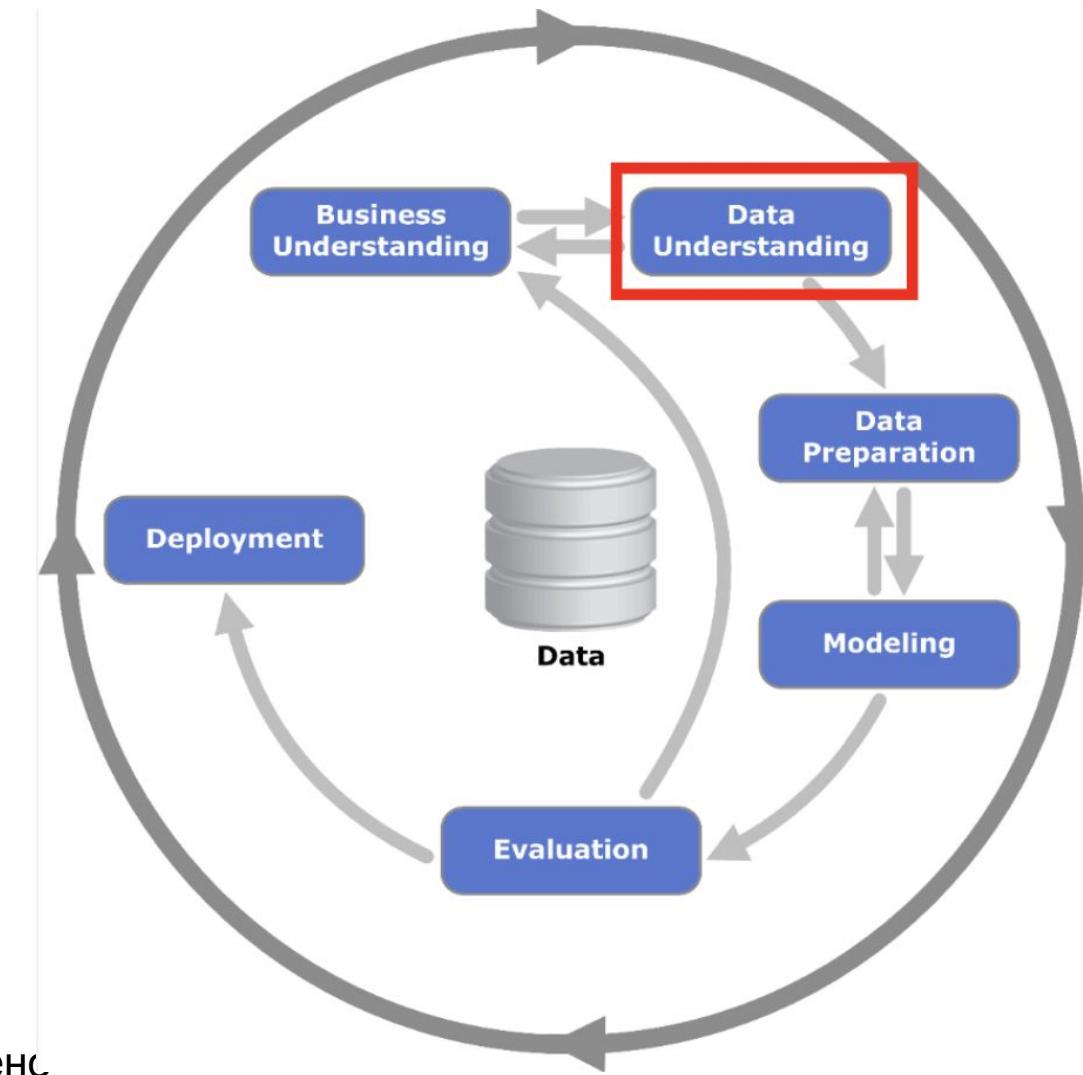
$\phi_i : D \rightarrow F_j$  - признак

Виды признаков:

- 1) Бинарные
- 2) Номинальные
- 3) Порядковые
- 4) Количественные

- Binary
- Categorical
- Ordinal
- Numerical

$F_j = \{true, false\}$   
 $F_j$  – конечно  
 $F_j$  – конечно упорядочено  
 $F_j = \mathbb{R}$



## Пример:

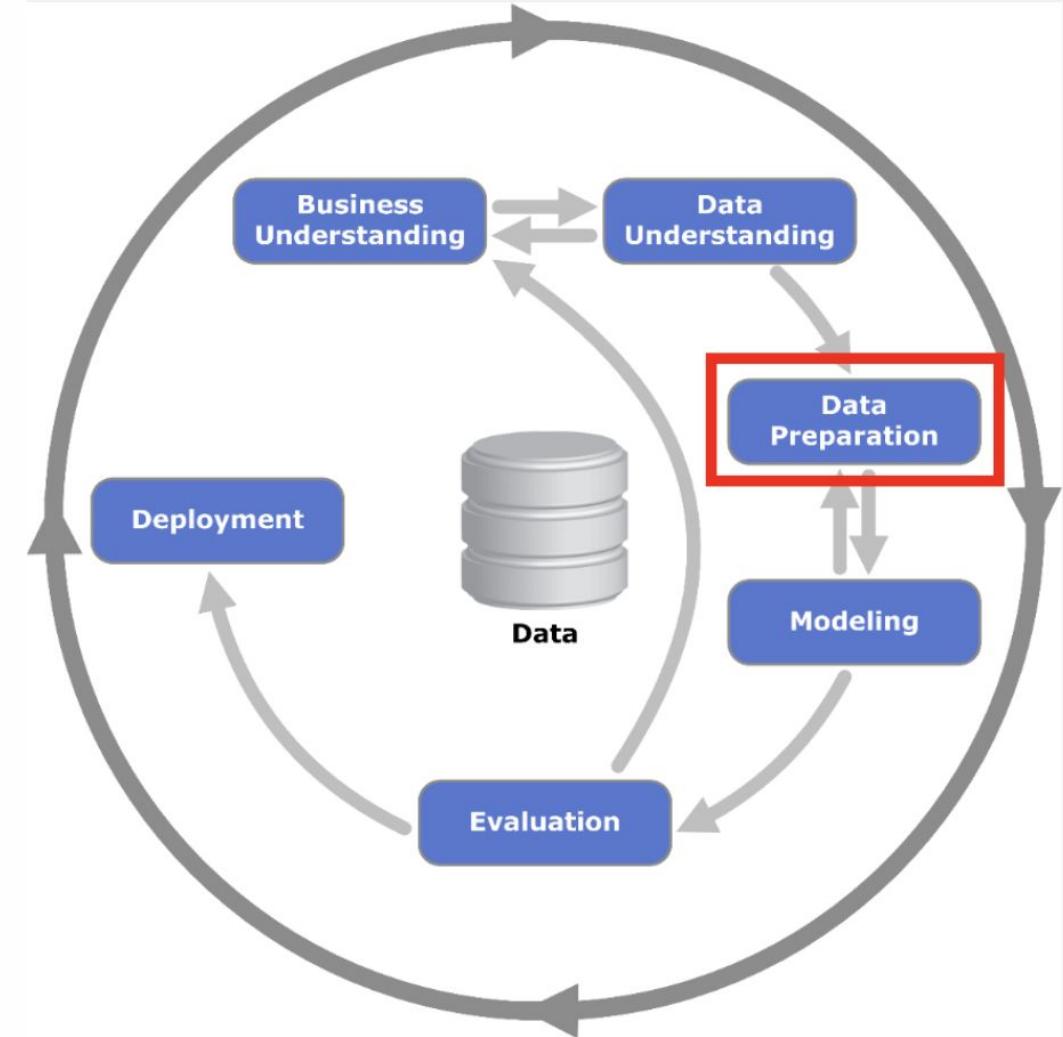
Задача: Необходимо спрогнозировать стоимость дома

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

- Удаление шума
- Заполнение отсутствующих значений
- Трансформация значений
- Генерация данных
- Выбор факторов
- Использование априорных знаний



# Создание модели (Modeling)

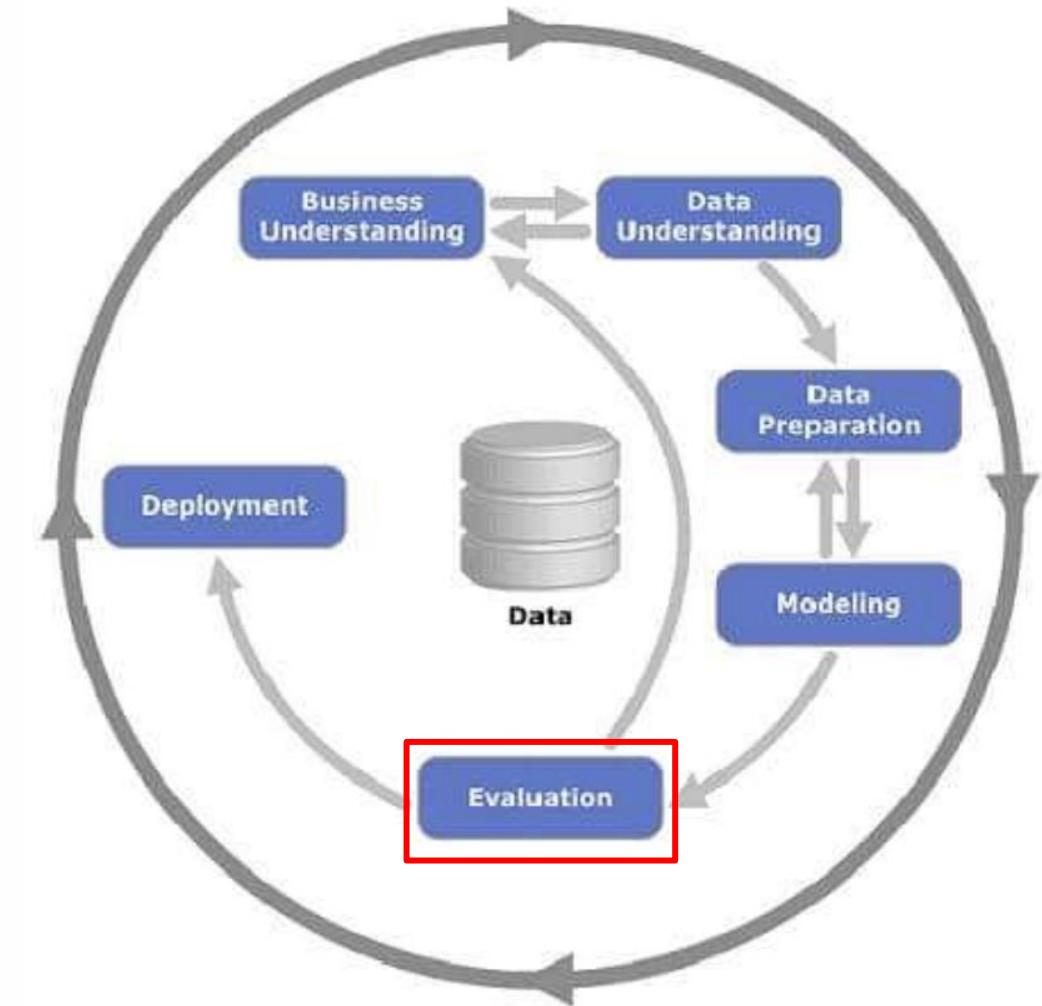
В зависимости от постановки задачи выбираются различные подходы к построению модели, описывающей свойства исследуемых объектов



# Оценка решения

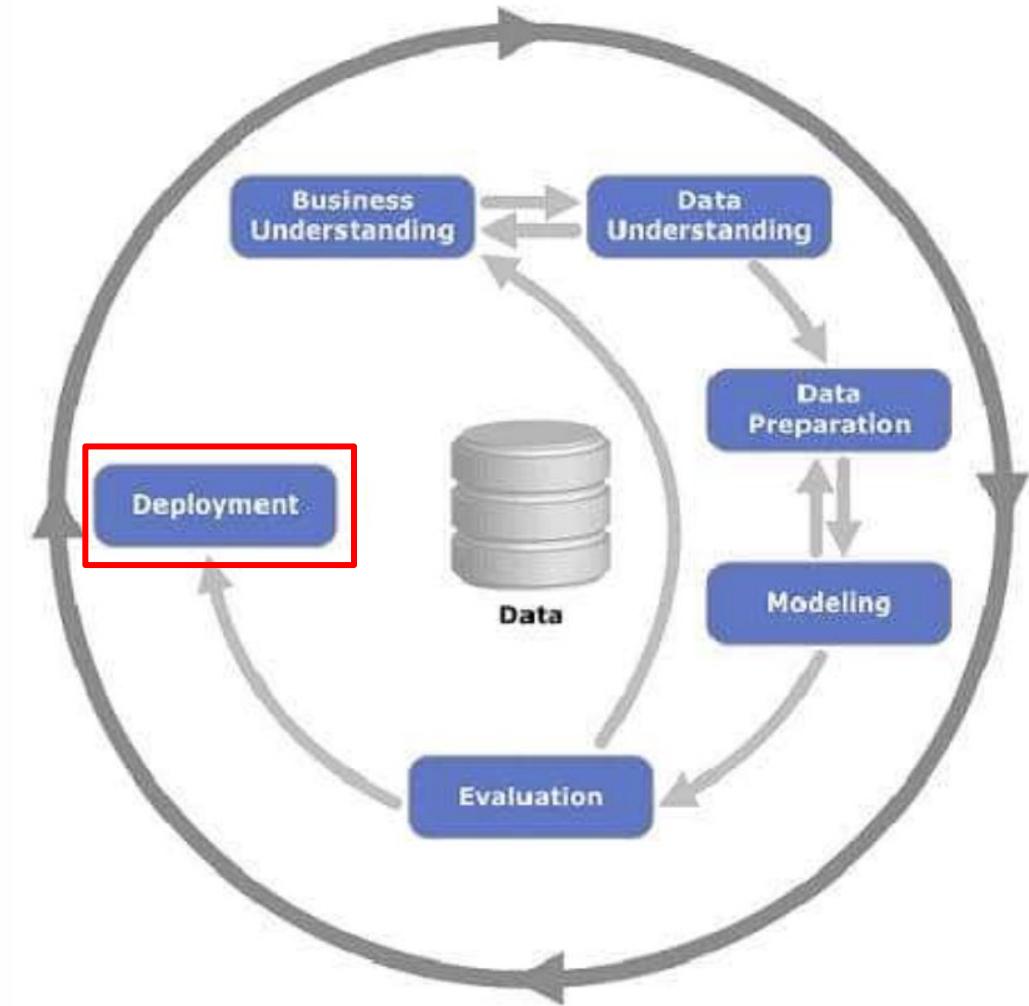
Оценка результатов с точки зрения достижения  
бизнес-целей

- Оценить результаты
- Сделать ревью процесса
- Определить следующие шаги



# Внедрение

- Запланировать развертывание
- Запланировать поддержку и мониторинг развернутого решения
- Сделать финальный отчет
- Сделать ревью проекта



---

**Окей, это конечно всё интересно. На зачем нам  
нужен C++???**



#024

# Окей, это конечно всё интересно. На зачем нам нужен C++???

## Оптимизация

Перепишите любой код с языка более высокого уровня на C++, чтобы программа работала быстрее. Так часто делают в сфере глубокого обучения и других алгоритмических областях, где важна скорость.

Подход может выглядеть так:

- пишу код на Python;
- заставляю работать нейросеть, которая решит мою проблему;
- переношу код на C++.

Но это не всегда приносит пользу: иногда быстрая разработка важнее производительности, или выгода от переноса кода на C++ совсем незначительна.

---

# **Окей, это конечно всё интересно. На зачем нам нужен C++???**

Чтобы понимать целостность картины.



---

## **Окей, это конечно всё интересно. На зачем нам нужен C++???**

Кто считает, что эти знания не нужны ML-щикам, то  
после этого семестра рекомендуем прочесть книгу

# Окей, это конечно всё интересно. На зачем нам нужен C++???

Кто считает, что эти знания не нужны ML-щикам, то после этого семестра рекомендуем прочесть книгу

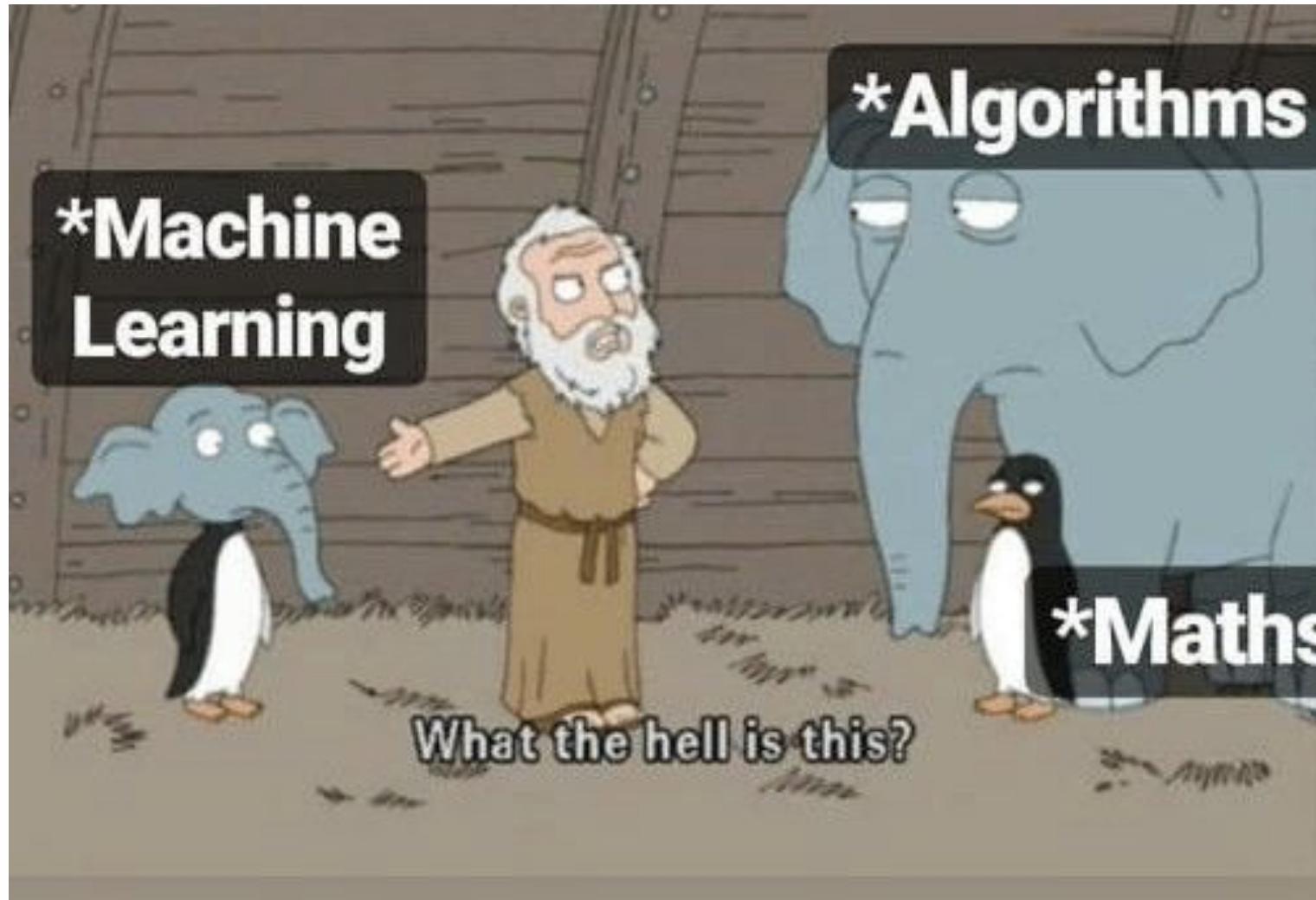




# Так что же такое ML?

#029

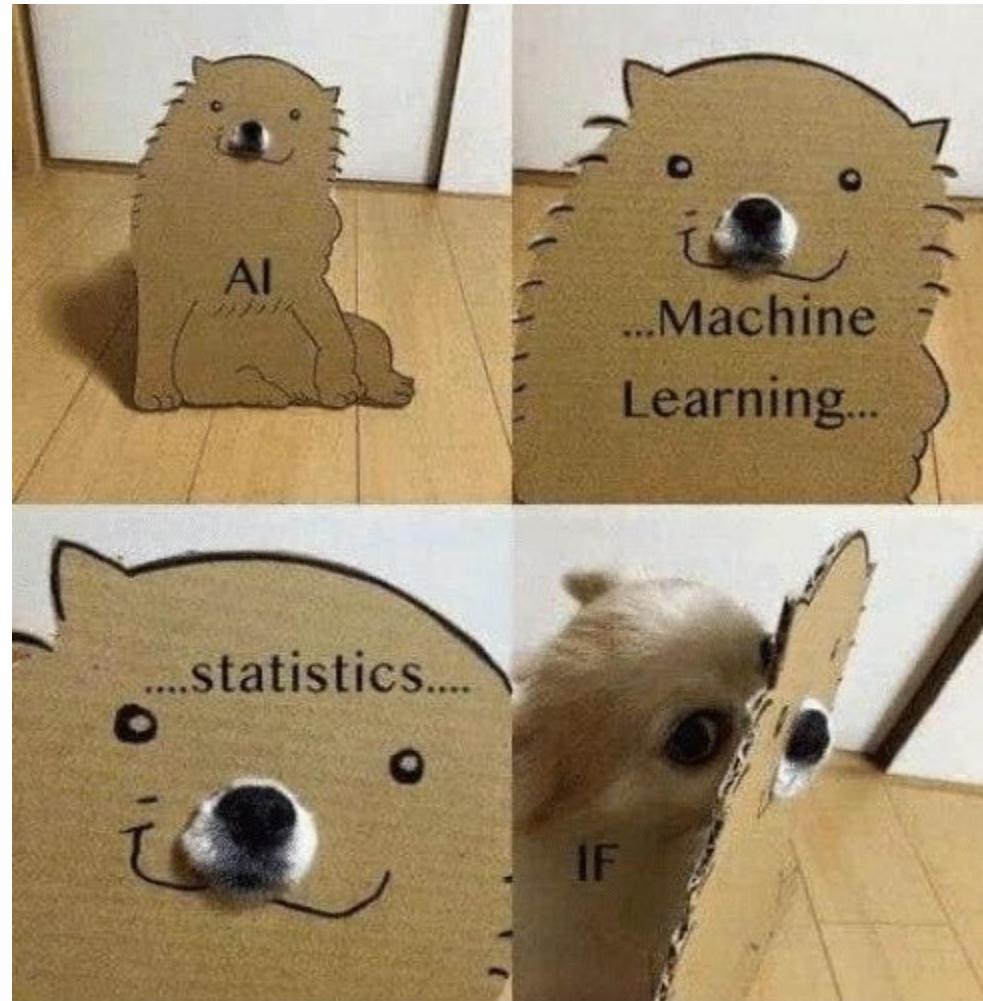
# Так что же такое ML?



#030

---

Или же?



#031

IF IF IF IF IF WE!

# 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.

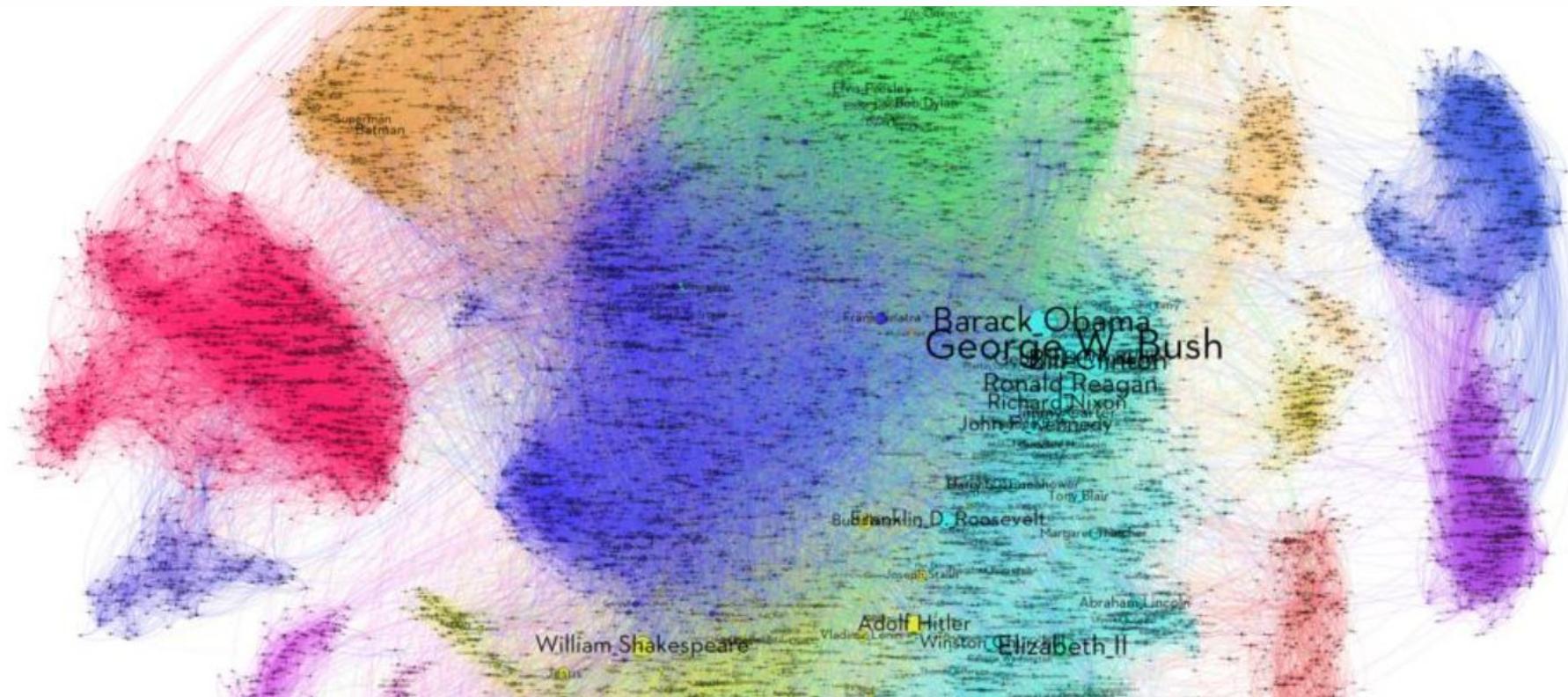
## 4) Обучение с подкреплением (**reinforcement learning**)

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

...

# Пример обучения без учителя

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



---

# Пример обучения без учителя

Поиск музыки одинакового жанра



#034

# Пример обучения без учителя

## Рекомендательные системы

КиноПоиск   Активировать промокод  Фильмы, сериалы, персоны   

Фильмы Сериалы  
С высоким рейтингом  
Российские Зарубежные  
Скрыть просмотренные

Жанры  
Со всеми жанрами

Страны  
Со всеми странами

Годы  
Со всеми годами

## Рекомендации

Персональные рекомендации создаются автоматически, они основаны на ваших оценках. Чтобы сделать рекомендации точнее, отмечайте просмотренные фильмы и сериалы и не стесняйтесь исполь... Читать полностью

Онлайн Все  
84 фильма 100 фильмов

По порядку

Рейтинг	Название	Год	Страна	Жанр	Просмотров	Действия
9.0	Игра престолов	2011–2019	США, Великобритания	фэнтези, драма	535 974	 Буду смотреть    
8.5	Джентльмены	2019	Великобритания, США	боевик, комедия	315 323	 Буду смотреть    
8.3	Зеленая книга	2018	США, Китай	комедия, драма	336 109	 Буду смотреть    

 Рекомендации

#035

# Пример обучения без учителя

Рекомендательные системы

Всё ли нормально???

#036

КиноПоиск   Активировать промокод  Фильмы, сериалы, персоны   

Фильмы Сериалы  
С высоким рейтингом  
Российские Зарубежные  
Скрыть просмотренные

Жанры  
Со всеми жанрами

Страны  
Со всеми странами

Годы  
Со всеми годами

## Рекомендации

Персональные рекомендации создаются автоматически, они основаны на ваших оценках. Чтобы сделать рекомендации точнее, отмечайте просмотренные фильмы и сериалы и не стесняйтесь исполь... Читать полностью

**Онлайн Все**  
84 фильма 100 фильмов

По порядку

Рейтинг	Название	Год	Страна	Жанр	Просмотров	Действия
9.0	Игра престолов	2011–2019	США, Великобритания	фэнтези, драма	535 974	    
8.5	Джентльмены	2019	Великобритания, США	боевик, комедия	315 323	    
8.3	Зеленая книга	2018	США, Китай	комедия, драма	336 109	    

**Рекомендации**



# 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  $\Theta$

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

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)

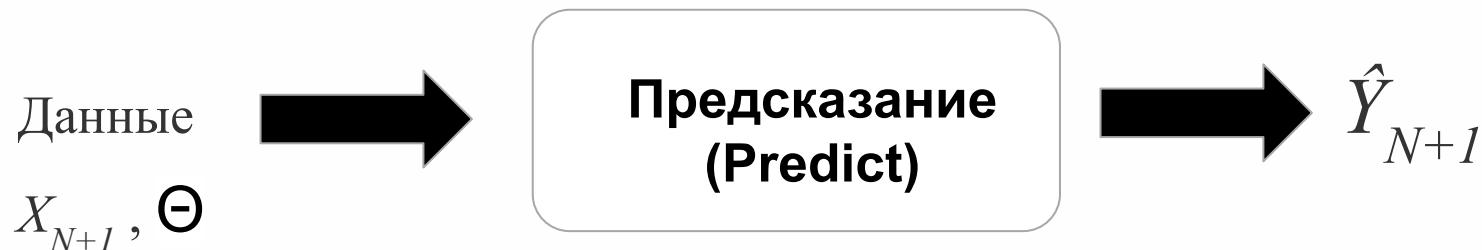
- $F_j$  - конечно упорядочено (ответы надо получить сразу на множестве объектов, после чего отсортировать их по значениям ответов)

# Обучение с учителем (обучения по прецедентам)

Этап обучения (train)



Этап применения (test)



# Обучение с учителем (обучения по прецедентам)

## Пример задачи классификации

Болеет ли человек коронавирусом? (бинарная классификация)



#040

# Обучение с учителем (обучения по прецедентам)

## Пример задачи классификации

Какой дорожный знак на изображении? (классификация на  $M$  непересекающихся классов)



#041

# Обучение с учителем (обучения по прецедентам)

## Пример задачи классификации

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



#042

# Обучение с учителем (обучения по прецедентам)

## Пример задачи классификации

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



# Обучение с учителем (обучения по прецедентам)

## Пример задачи классификации

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



---

## Обучение с учителем (обучения по прецедентам)

Пример задачи восстановления регрессии

Предсказание курса валют



# Обучение с учителем (обучения по прецедентам)

## Пример задачи восстановления регрессии

Предсказание уровня зарплаты по  
резюме

#046

```
/* resume.css
   REZUME IT-SHNIKA

#Objective {
    Employment: Internship, Part-time;
}

#Skills .WebDesign {
    Language: HTML, CSS, JavaScript, MySQL;
    Software: Illustrator, Photoshop, Flash;
}

#Skills .OtherMedia {
    Software: After-Effects, Premiere-Pro, Soundbooth,
    Media-Encoder;
}

#Work-Experience {
    Employee: PCKIZ;
    Position: Summer-Intern;
    Responsibility: Customizing-Google-Maps, Marking-Up-
Facebook-Page-Canvas-Tab, Managing-Social-Media;
}

#Education {
    Major: Multimedia-Arts Web-Design;
    Class-Standing: Senior;
}

#ContactInfo {
    Name: Shanning-Wan;
    Email: shanning@makewan.com;
    Skype-Username: shannning;
}
```

# Обучение с учителем (обучения по прецедентам)

## Пример: задачи ранжирования



Пример: задачи ранжирования



Все

Картинки

Новости

Видео

Покупки

Ещё

Настройки

Инструменты

Результатов: примерно 1 660 000 (0,49 сек.)

[ru.coursera.org › lecture › data-analysis-applications](#) ▾

### [Задача ранжирования - Рекомендации и ранжирование ...](#)

На их **примере** вы узнаете, как извлекать признаки из разнородных данных, какие при этом возникают проблемы и как их решать. Вы научитесь сводить ...

[neerc.ifmo.ru › wiki › title=Ранжирование](#) ▾

### [Ранжирование — Викиконспекты](#)

**Ранжирование** (англ. learning to rank) — это класс задач машинного обучения с ...

Линейная модель ранжирования: ... Пример вычисления DCG и nDCG:.

[edu.mmcs.sfedu.ru › mod › resource › view](#) ▾ PDF

### [Машинное обучение Ранжирование](#)

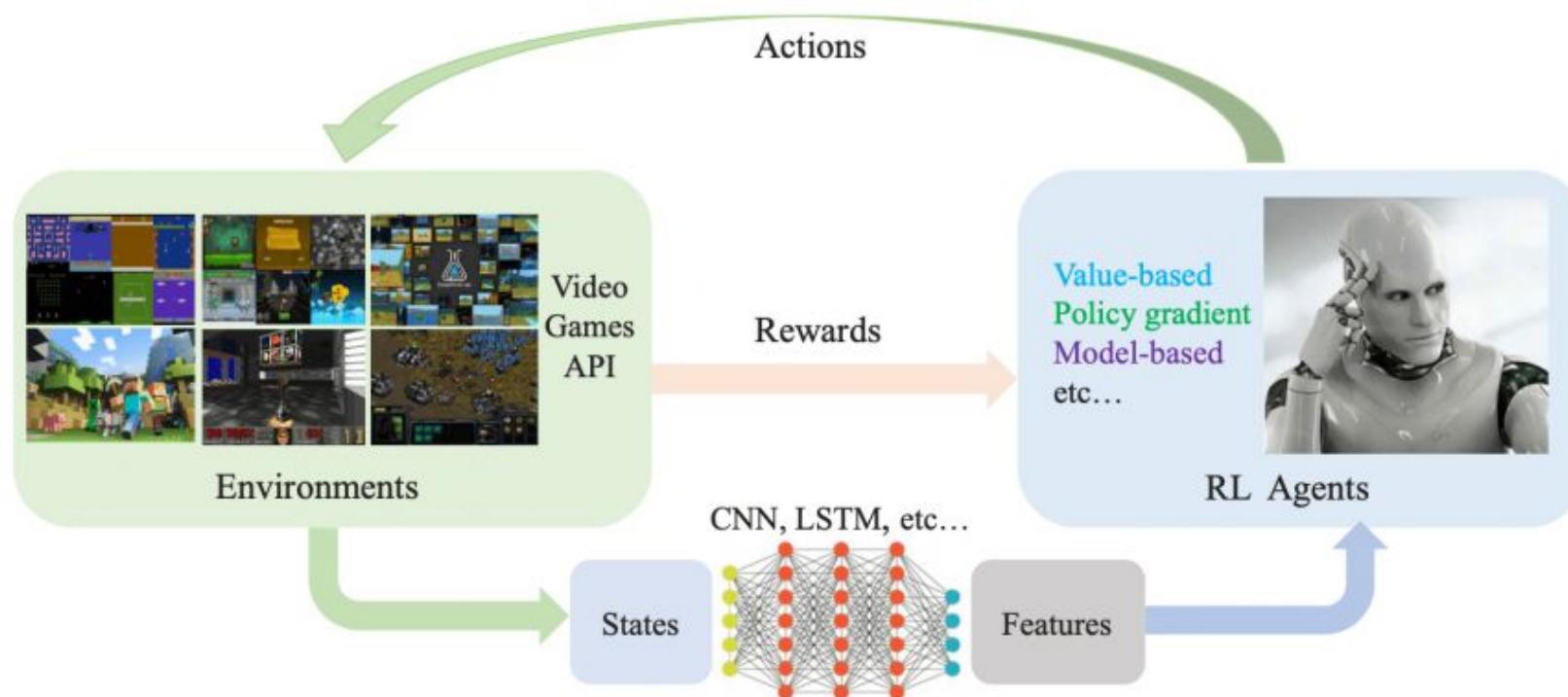
Оценки качества. ◦ Подходы к решению задачи. – поточечный ... 9. Пример вычисления nDCG ... Сведем задачу ранжирования к задаче предсказания.

[www.hse.ru › data › 2012/06/20 › Алгоритмы ранжирования и их п...](#) ▾ PDF

### [Алгоритмы ранжирования и их применение в задачах ...](#)

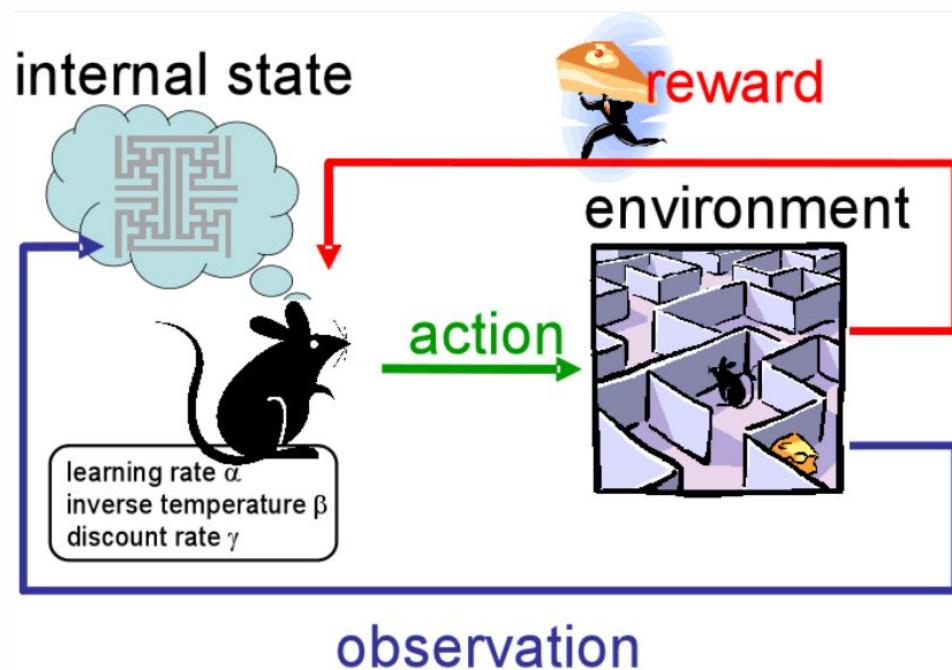
20 июн. 2012 г. - системах. Задачи работы: 1) Выявление существенных факторов, влияющих на ранжирование. 2) Кластеризация схожих запросов.

# Обучение с подкреплением (reinforcement learning)



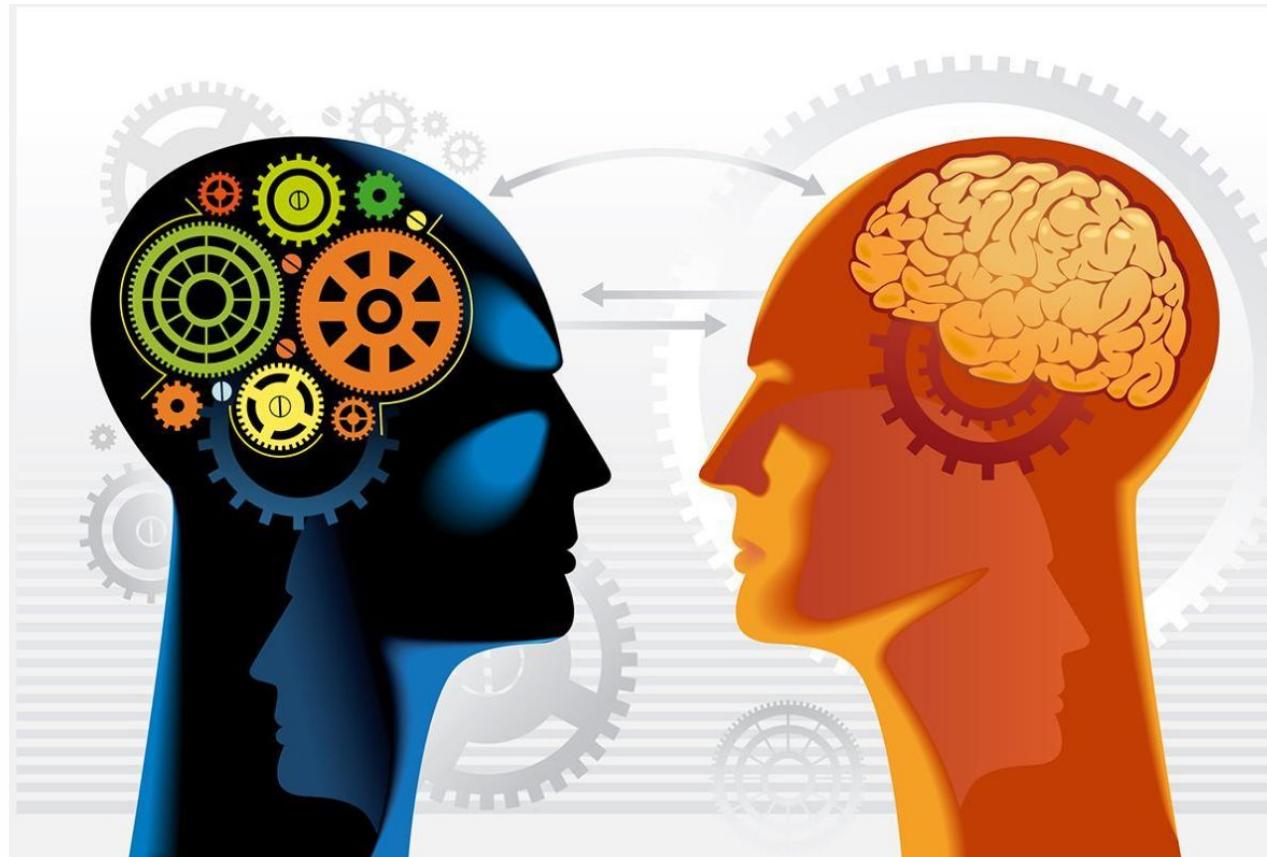
# Обучение с подкреплением (reinforcement learning)

Пример: игры atari



# Обучение с подкреплением (reinforcement learning)

Пример: создание чатбота



---

## Инструменты



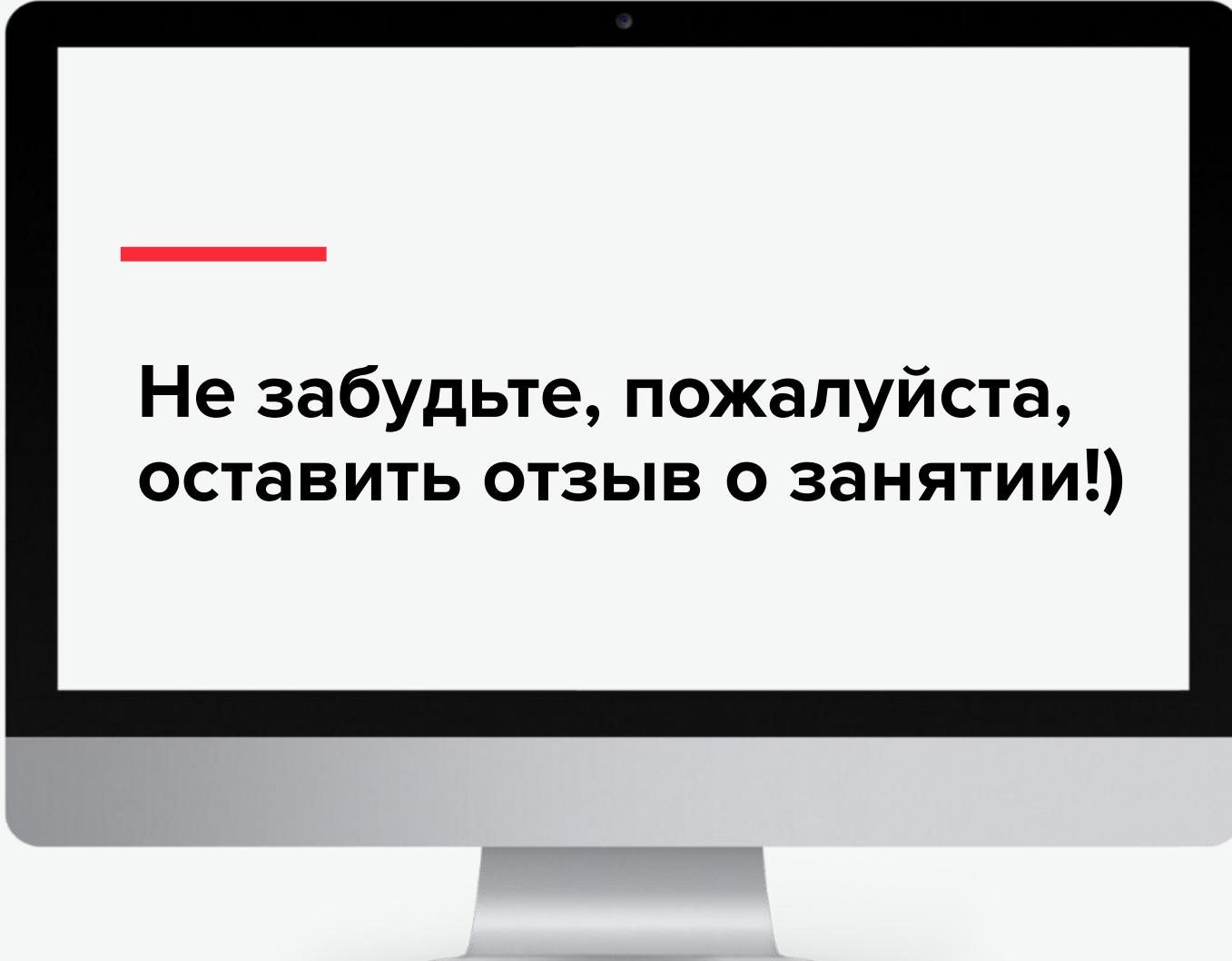
# Инструменты



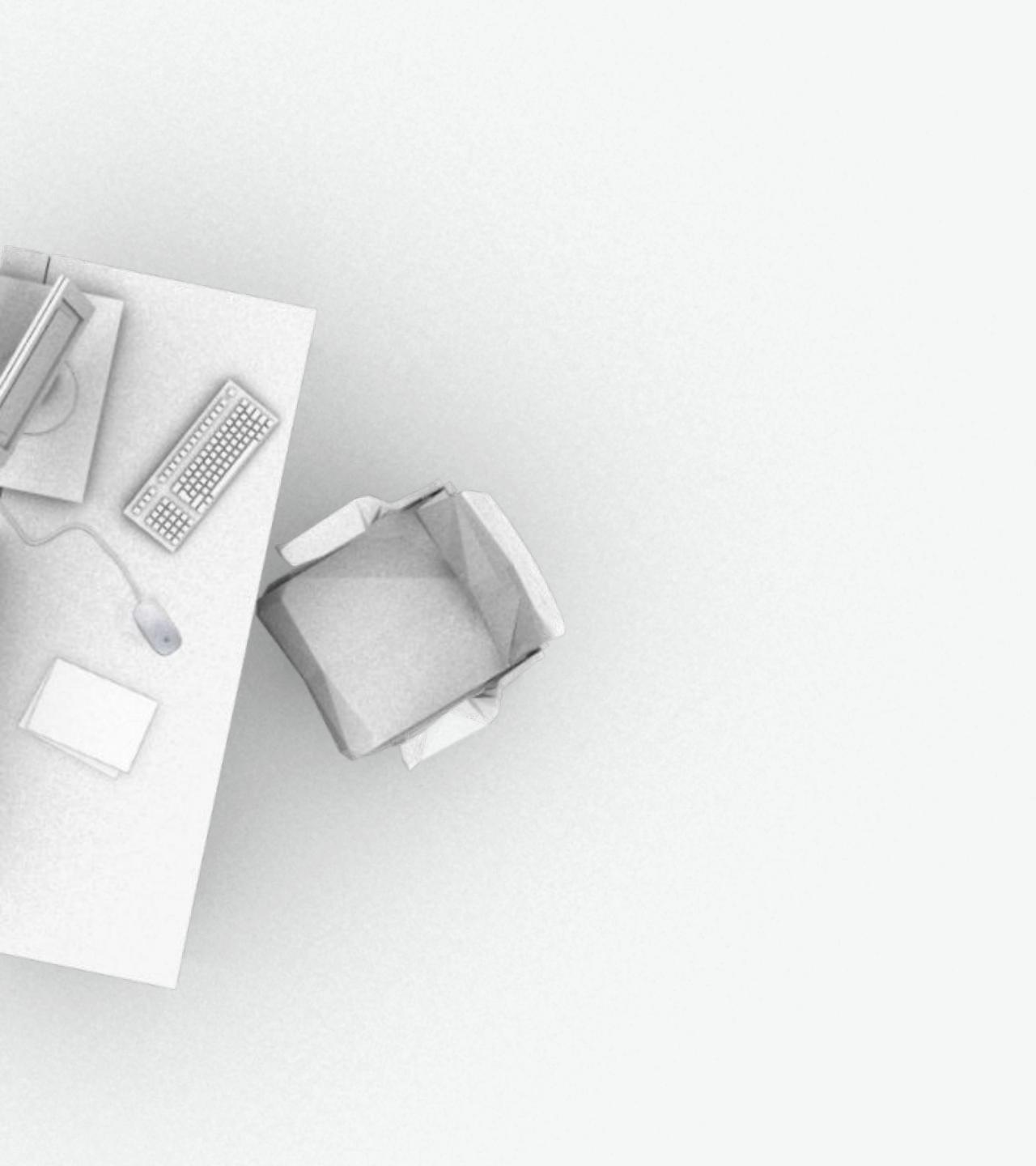
#052

# Python инструменты





**Не забудьте, пожалуйста,  
оставить отзыв о занятии!)**



# **Введение в машинное обучение**

Практическая часть