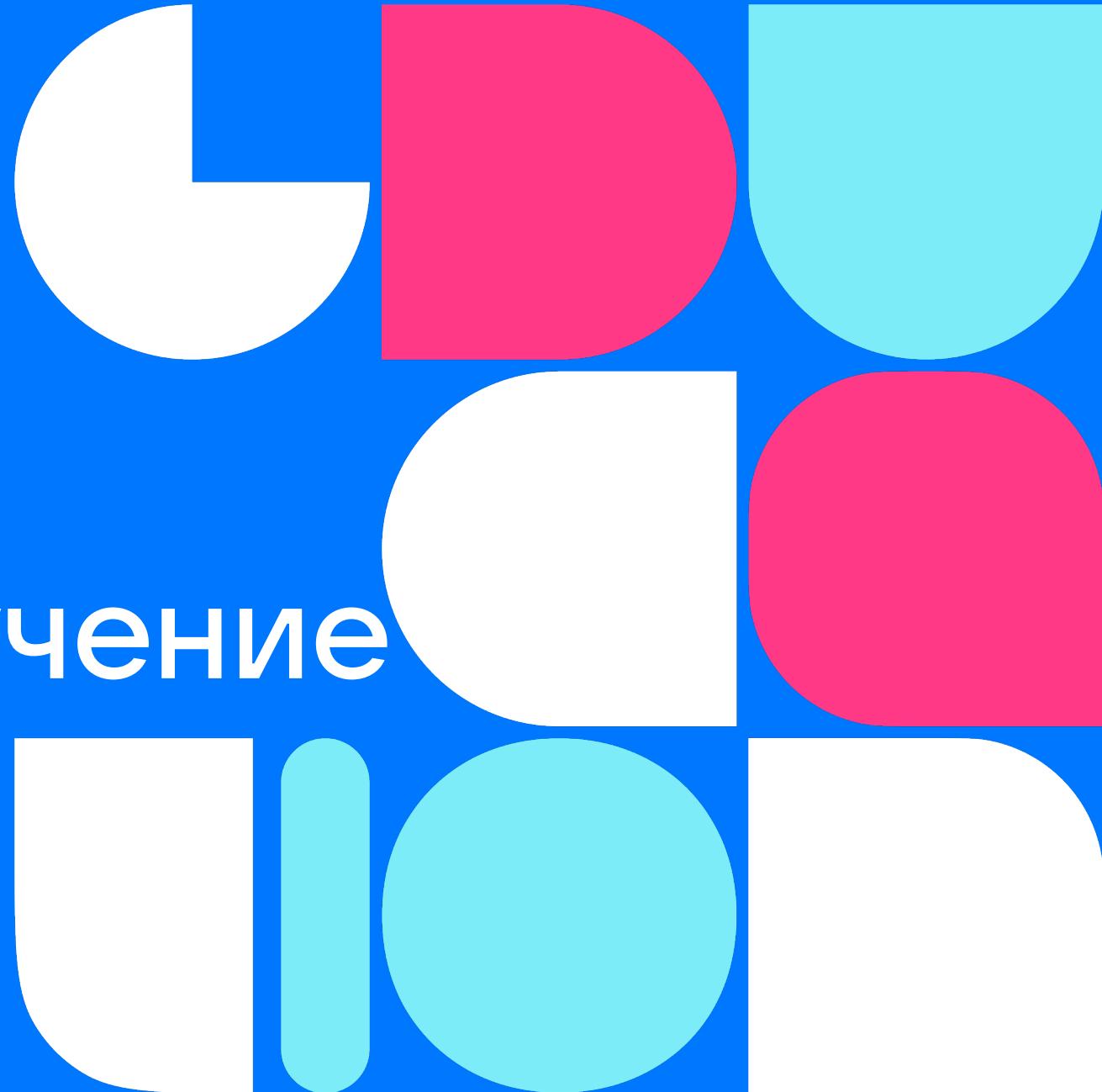




Введение в машинальное обучение

Лекция 1. Введение в анализ
данных и машинное обучение

Журавлёв Вадим



Не забываем отмечаться
на портале и оставлять
отзывы!

The screenshot shows a website interface for '@технопарк'. At the top, there is a navigation bar with links: Блоги, Люди, Программа, Выпуски, Расписание (selected), Вакансии, and a user profile for 'Вадим Журавлёв'. Below the navigation is a search bar and a download icon.

The main content area displays a list of events for a course titled 'Основы машинного обучения' (Machine Learning Basics). The events are listed by date and time:

- 30 сентября 18:00 — 21:00 среда: Основы машинного обучения. Смешанное занятие 1. Уточняется. ML-11
- 7 октября 18:00 — 21:00 среда: Основы машинного обучения. Смешанное занятие 2. Уточняется. ML-11
- 14 октября 18:00 — 21:00 среда: Основы машинного обучения. Смешанное занятие 3. Уточняется. ML-11
- 21 октября 18:00 — 21:00 среда: Основы машинного обучения. Смешанное занятие 4. Уточняется. ML-11
- 28 октября 18:00 — 21:00 среда: Основы машинного обучения. Смешанное занятие 5. Уточняется. ML-11
- 3 ноября 18:00 — 21:00 вторник: Основы машинного обучения. Смешанное занятие 6. Уточняется. ML-11
- 11 ноября 18:00 — 21:00 среда: Основы машинного обучения. Смешанное занятие 7. Уточняется. ML-11

To the right of the event list is a calendar for the months of September, October, and November. The days of the week are labeled in Russian: Пн, Вт, Ср, Чт, Пт, Сб, Вс. Specific dates are highlighted in orange, such as 19, 20, 21, 28, 29, 30 in September, and 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30 in October.

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 интегралы)

3

Теория вероятностей и
математическая статистика
(понятия вероятности,
условной вероятности,
функции распределения)

4

Python (хотя бы чуть-чуть)

Преподаватели:



**Вадим
Журавлёв**

Руководитель группы
разработки моделей
ранжирования, VK



**Александр
Мамаев**

Руководитель группы
Группа дата-аналитики, VK



**Максим
Кулаев**

Руководитель группы
разработки CRM
моделей, VK



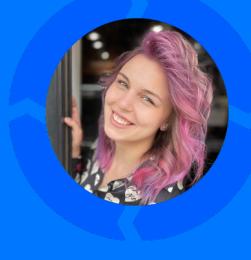
**Владислав
Ефимов**

Технический лидер, VK



**Константин
Шематоров**

Руководитель группы
Группа Autogen, VK



**Ангелина
Ярошенко**

Программист-исследователь,
VK

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. Коллоквиум
8. Ансамбли моделей
9. Работа с текстовыми данными II
10. Рекомендательные системы
11. Работа с гео-данными
12. CRM модели: response, look-a-like, uplift, контроль качества данных и мониторинг моделей
13. Введение в SQL
14. Экзамен

Выполнение заданий

1

Задания будут выкладываться на портал

2

Решения тоже нужно заливать на портал в виде ссылки на google colab

Домашние задания

Оценка за курс = оценка домашних заданий (0.6) + коллоквиум (0.2) + экзамен (0.2)

Правила сдачи ДЗ:



Экзамен

Что входит в экзамены?



Будет необходимо приготовить
презентацию-отчет по всем ДЗ.

На одно ДЗ – 1-2 слайда



Ответы на вопросы по курсу

Введение в машинное обучение и анализ данных

Лекция 1



Содержание лекции



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 VK (ex. Mail.Ru Group):

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

Технопарк, Техносфера, Технотрек, Техноатом, Технополис

Machine Learning

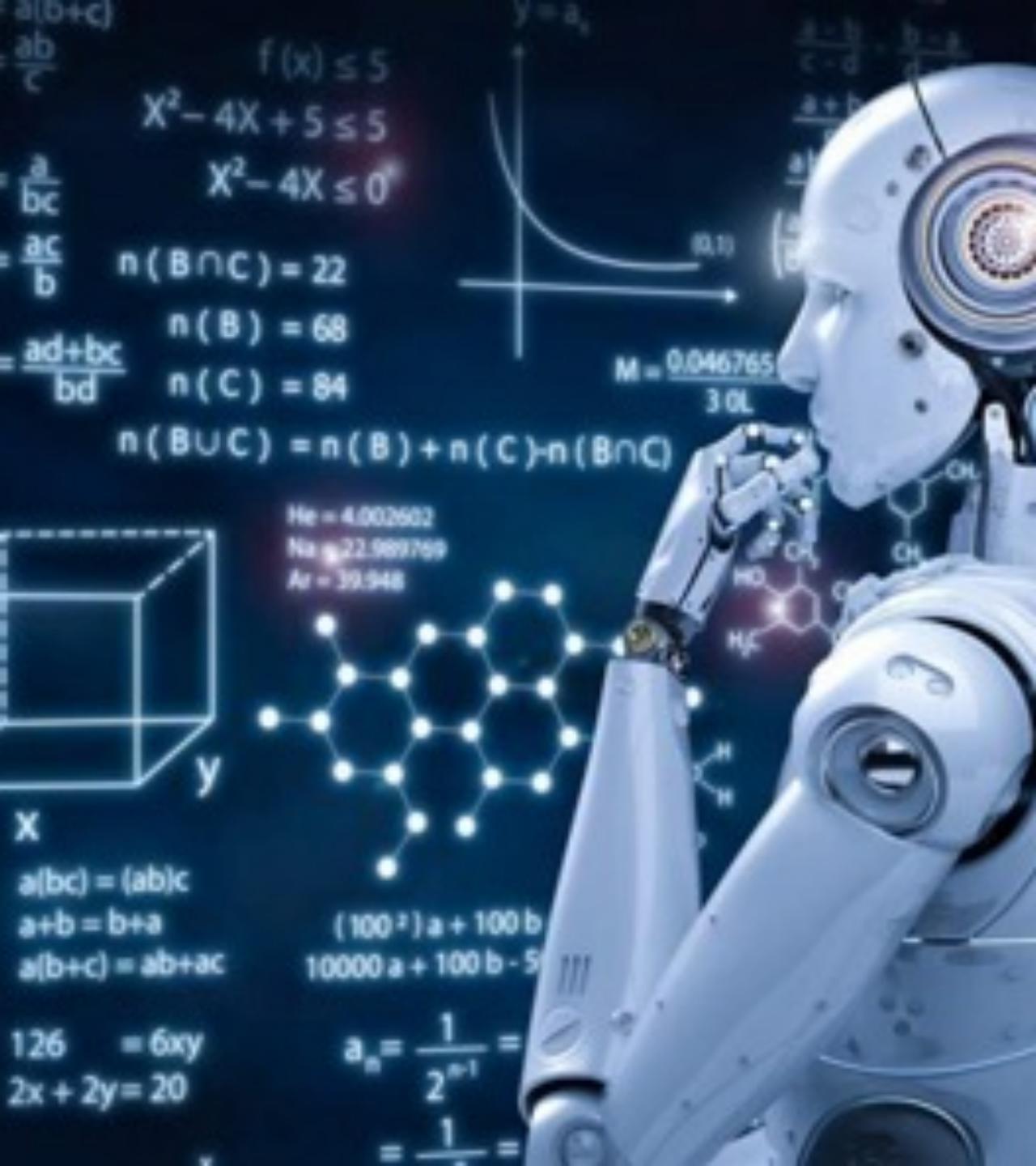


Машинное обучение (Machine Learning)

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



education



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.



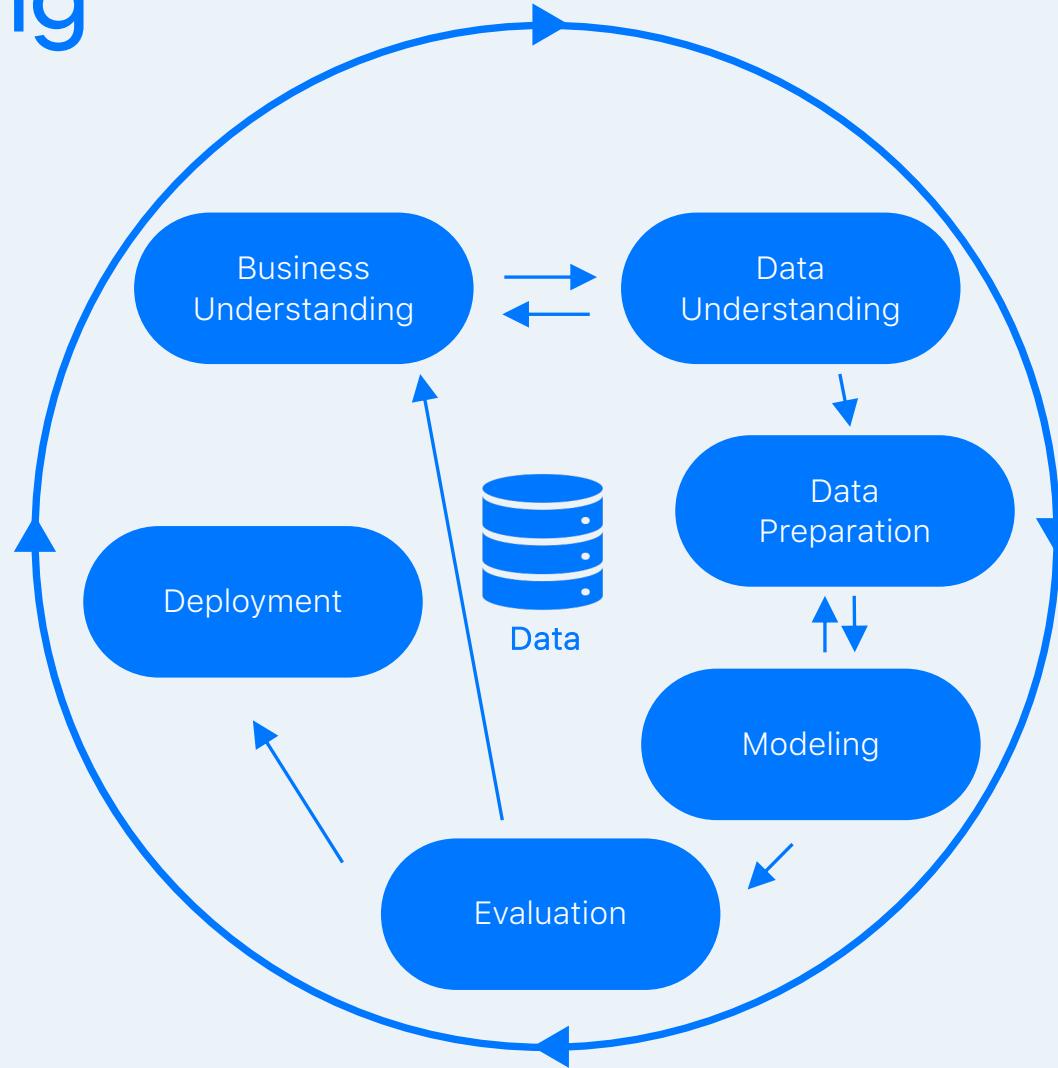
A photograph of a woman in a blue and white swimsuit teaching a young girl in a pink and blue swimsuit how to swim in a blue-tiled pool. The woman is holding the girl's arms and guiding her through the water. In the background, there is a white van parked behind a chain-link fence, and palm trees are visible under a clear sky.

Обучение модели

Data Scientist

Внедрение модели

CRISP-DM: Cross Industry Standard Process for Data Mining

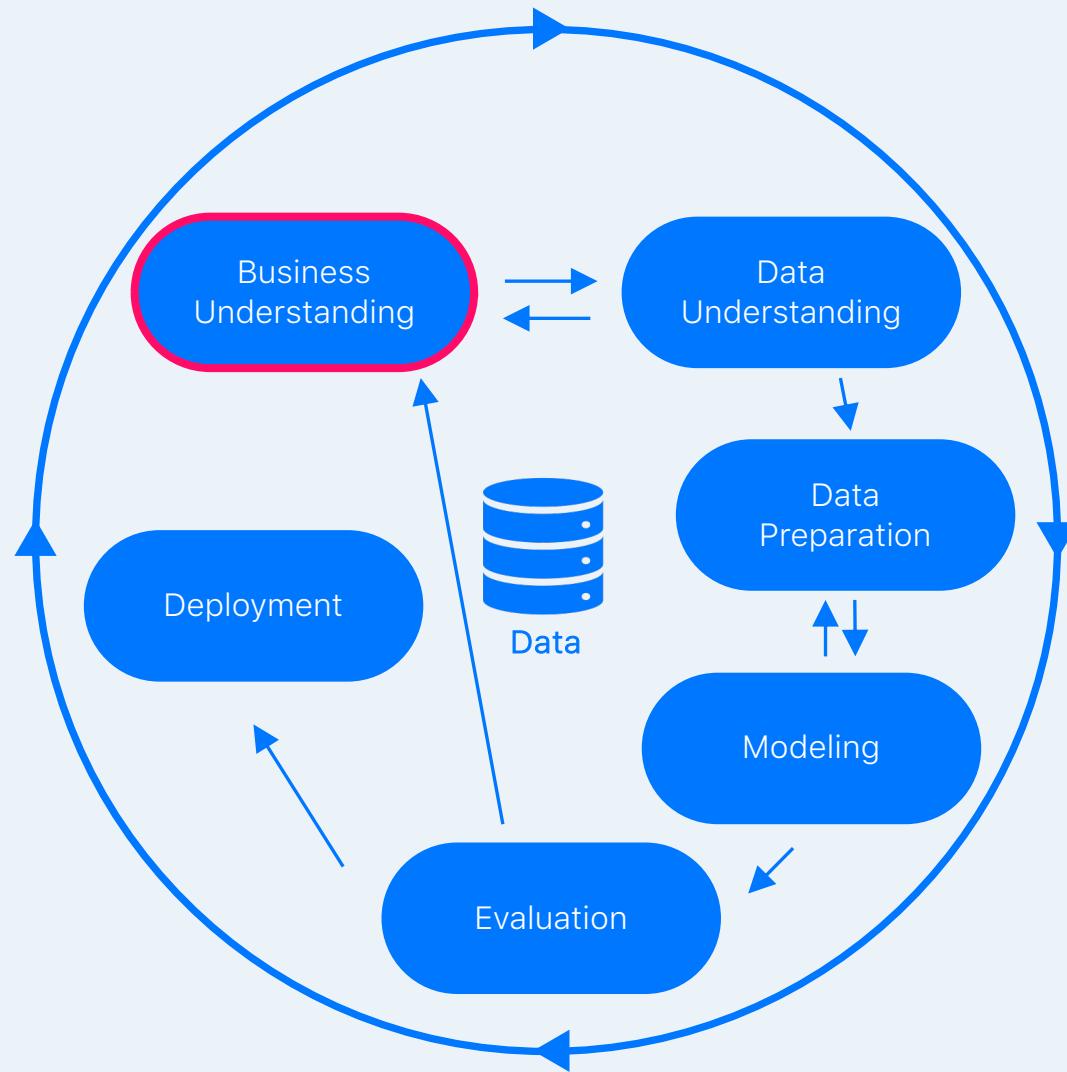


Бизнес-анализ:

- 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)

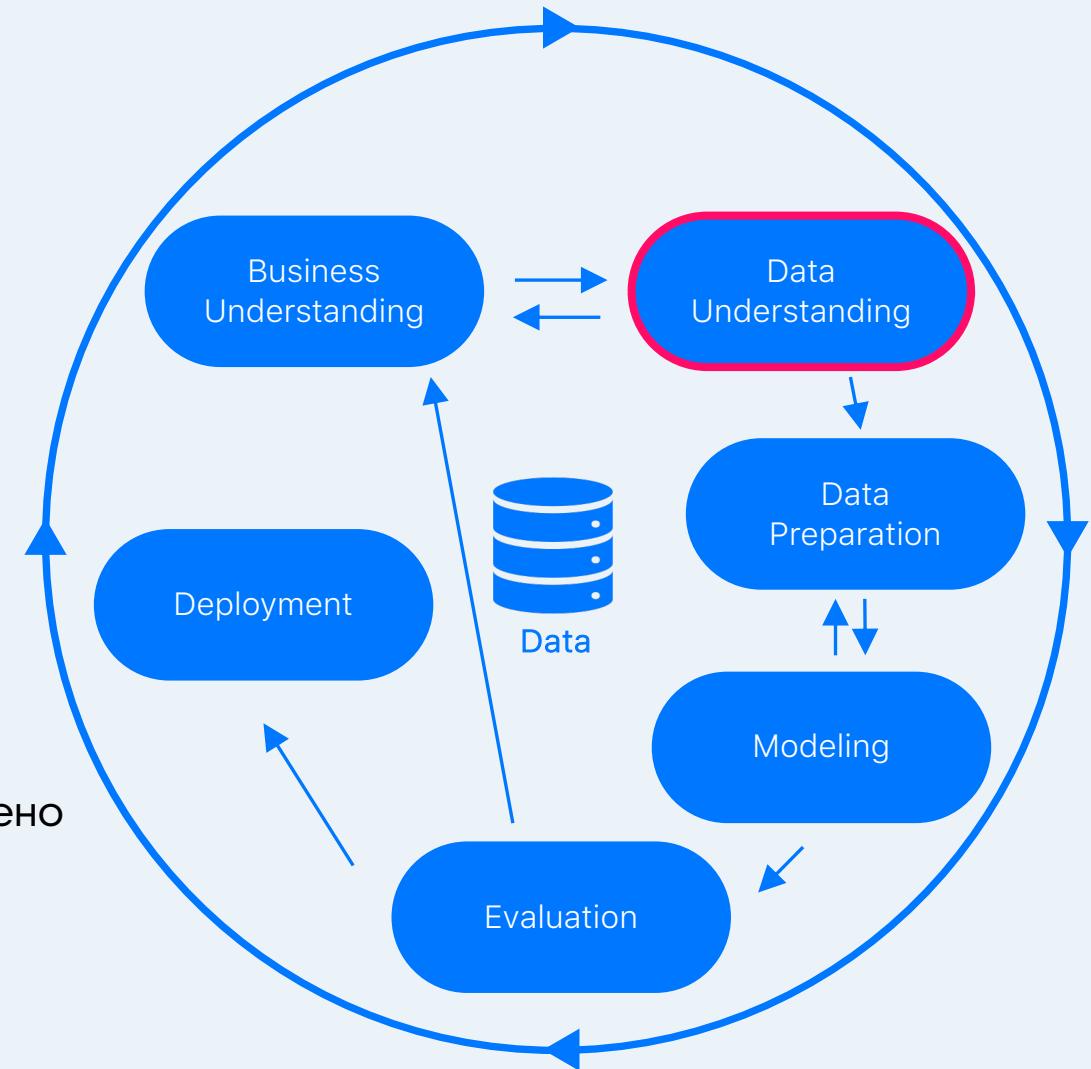
D – множество объектов (Data set)

$d \in D$ – обучающий объект

$\phi_i: D \rightarrow F_j$ - признак

Виды признаков:

1) Бинарные	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:

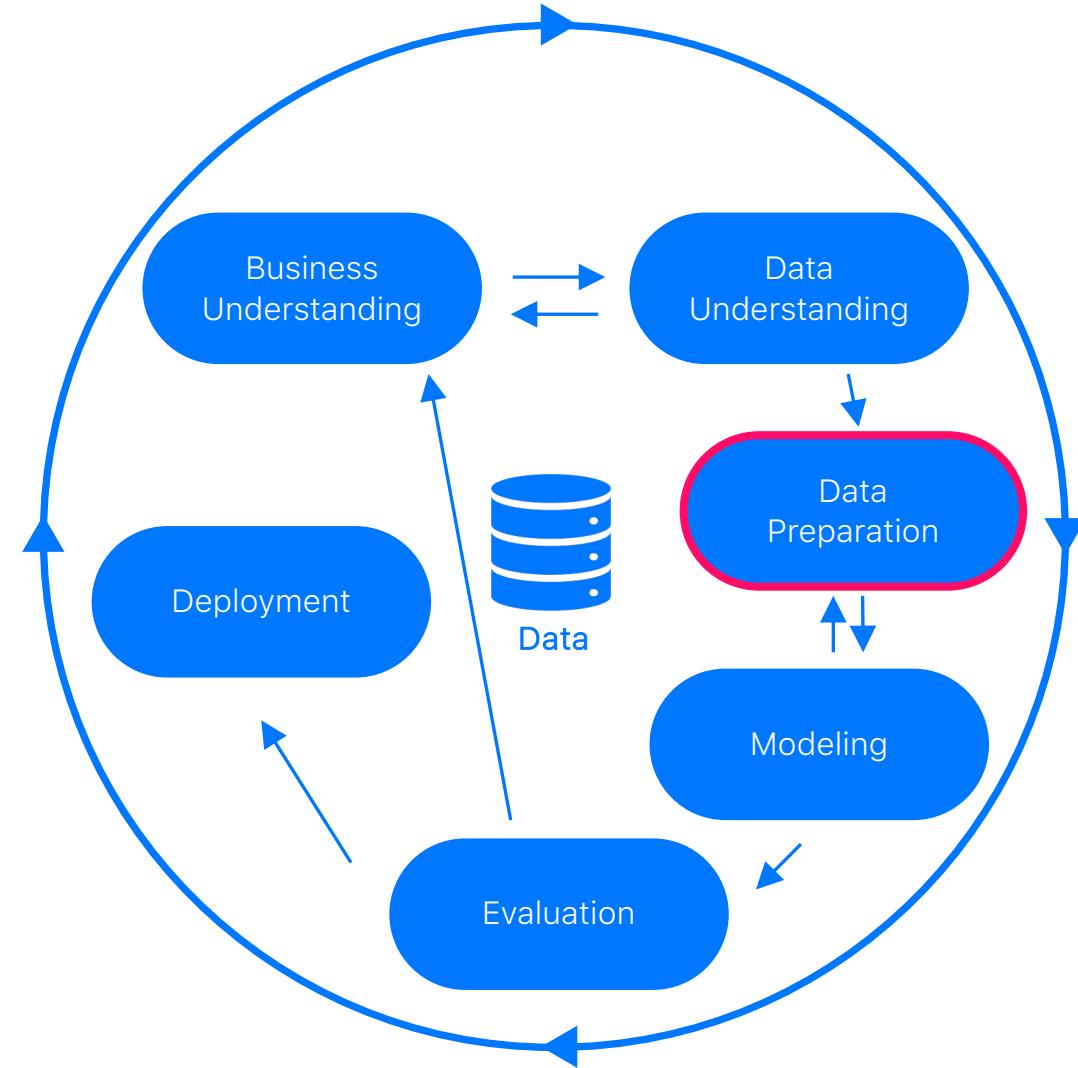
Задача: Необходимо спрогнозировать стоимость дома

Признаки, характеризующие стоимость жилья:

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

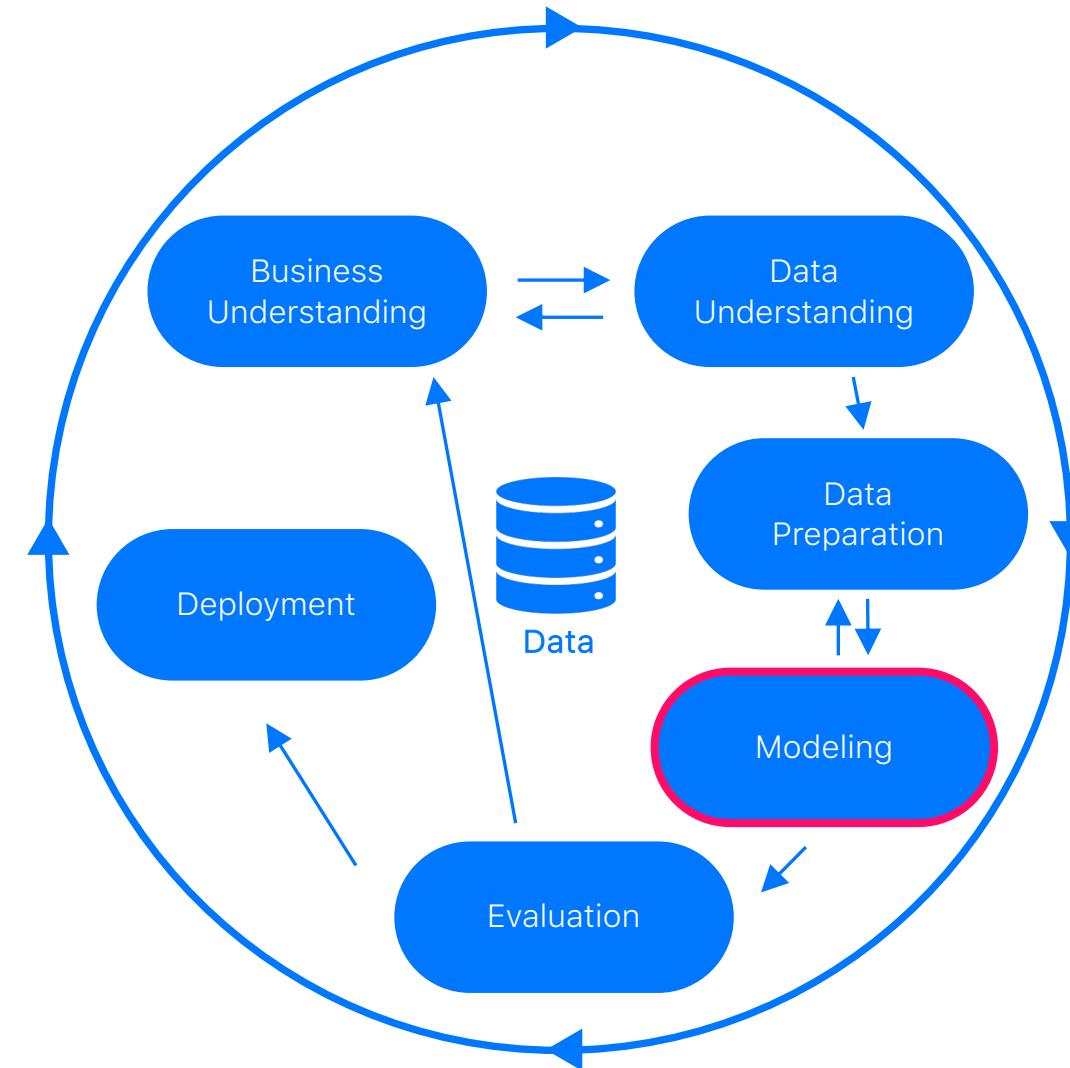
Подготовка данных

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



Создание модели (Modeling)

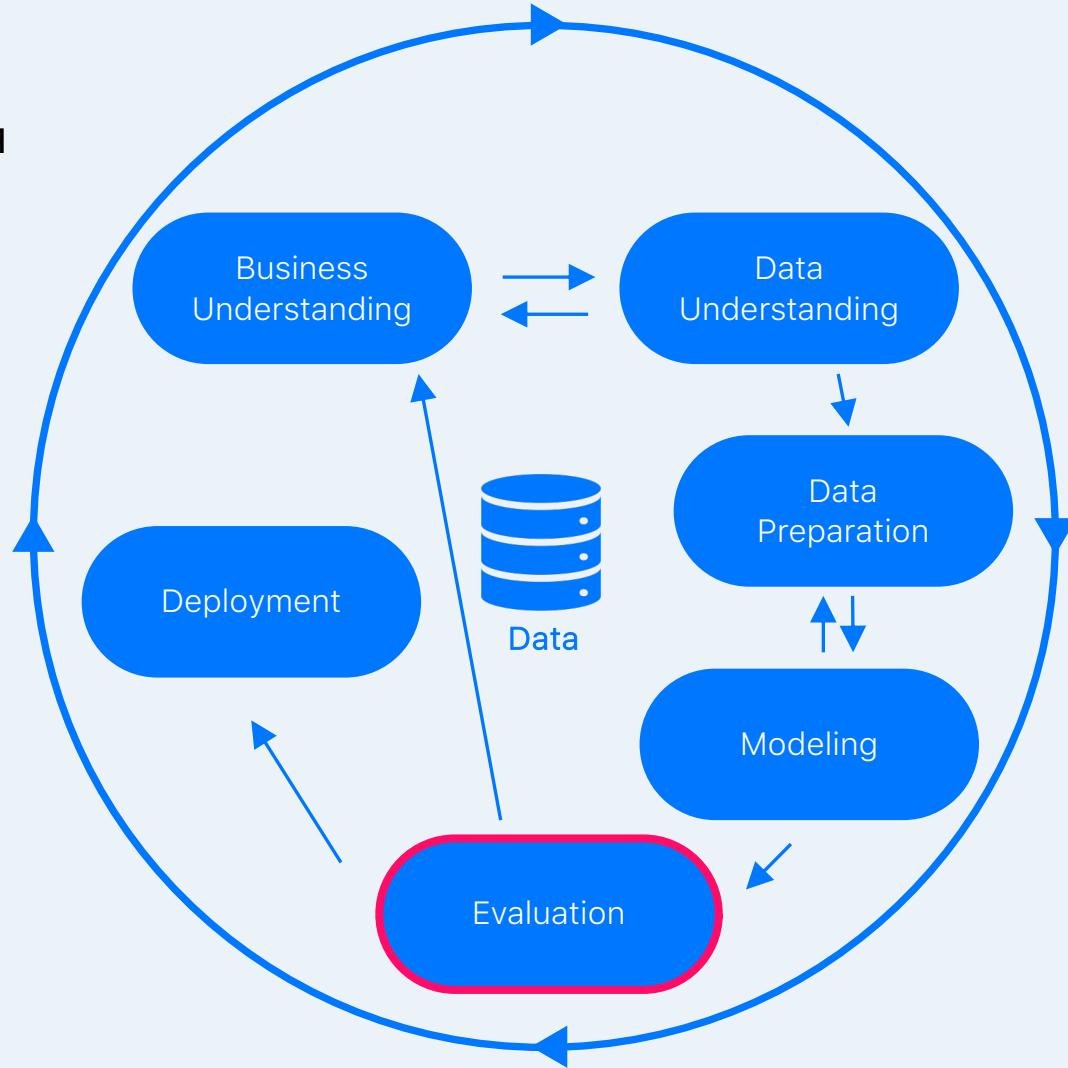
В зависимости от постановки задачи выбираются различные подходы к построению модели, описывающей свойства исследуемых объектов



Оценка решения

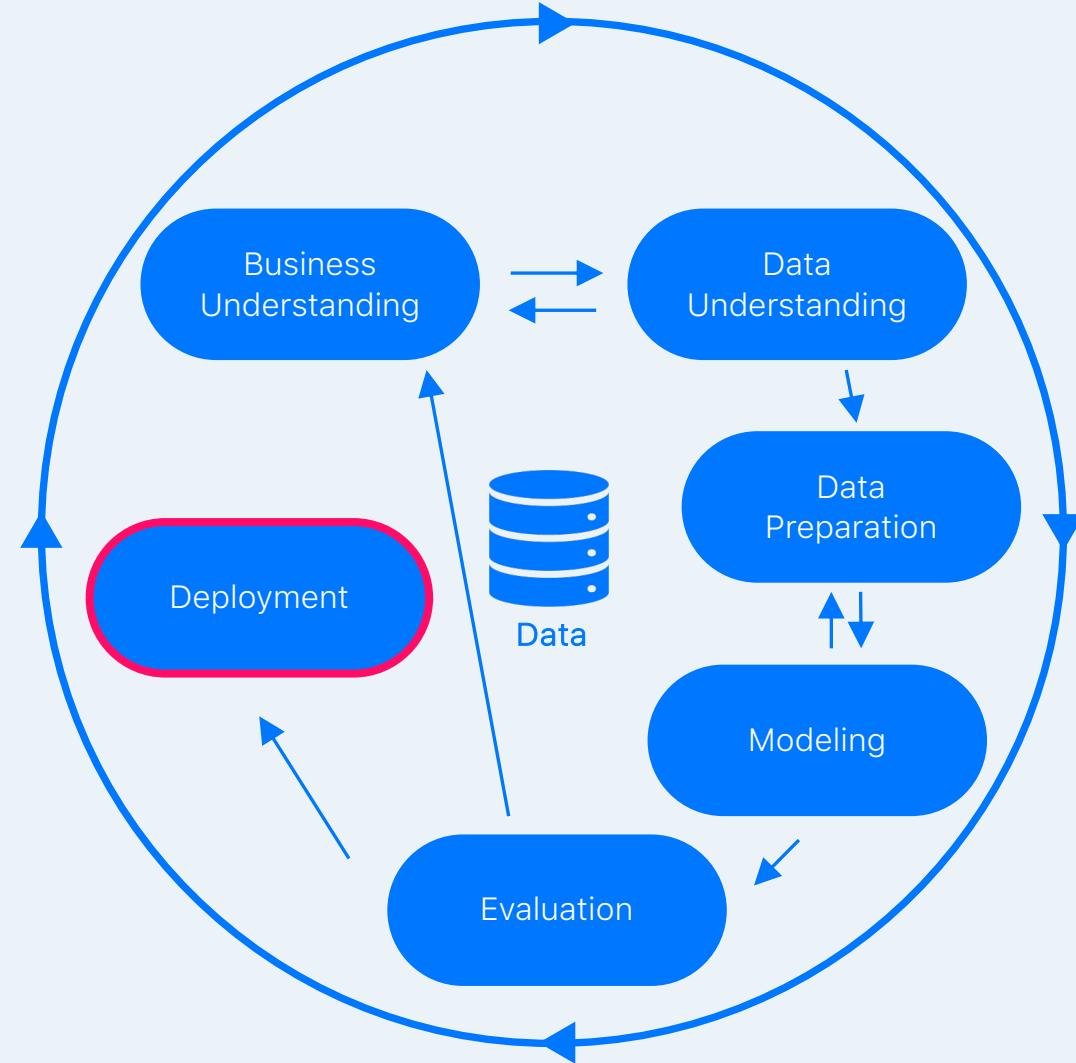
Оценка результатов с точки зрения достижения
бизнес-целей

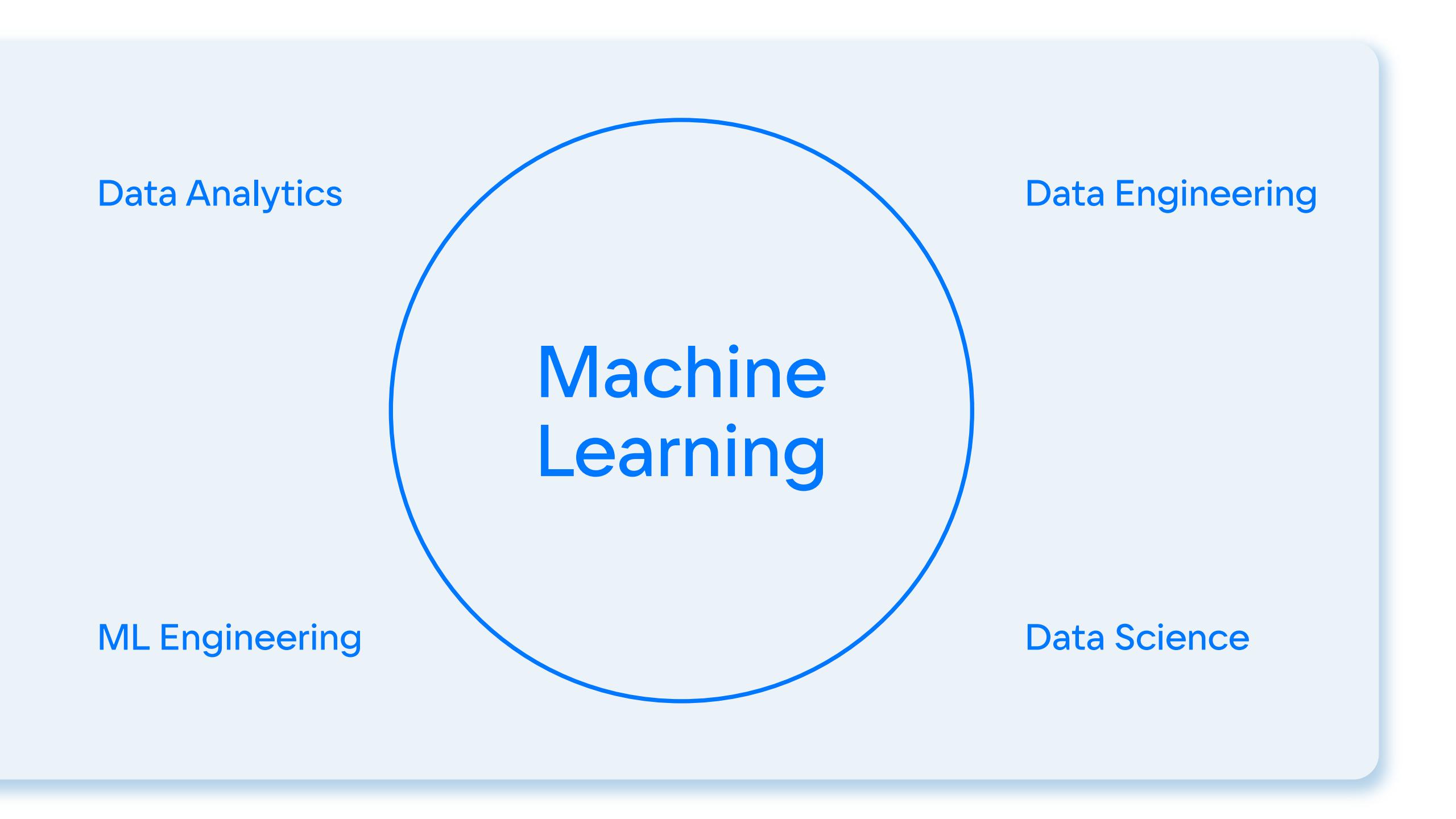
- Оценить результаты
- Сделать ревью процесса
- Определить следующие шаги



Внедрение

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





Data Engineering

Machine Learning

ML Engineering

Data Science

Data Analytics

Data Analytics

отчёты
дашборды

Data Engineering

почистить данные
подготовить фичи

ML Engineering

вывести в прод
оптимизировать

Machine Learning

Data Science

обучить модельки
исследовать гипотезы

Основные типы задач

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.

1

Частичное обучение (semi-supervised learning)

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

3

Обучение без учителя (unsupervised learning)

Ответы не задаются, и требуется искать зависимости между объектами.

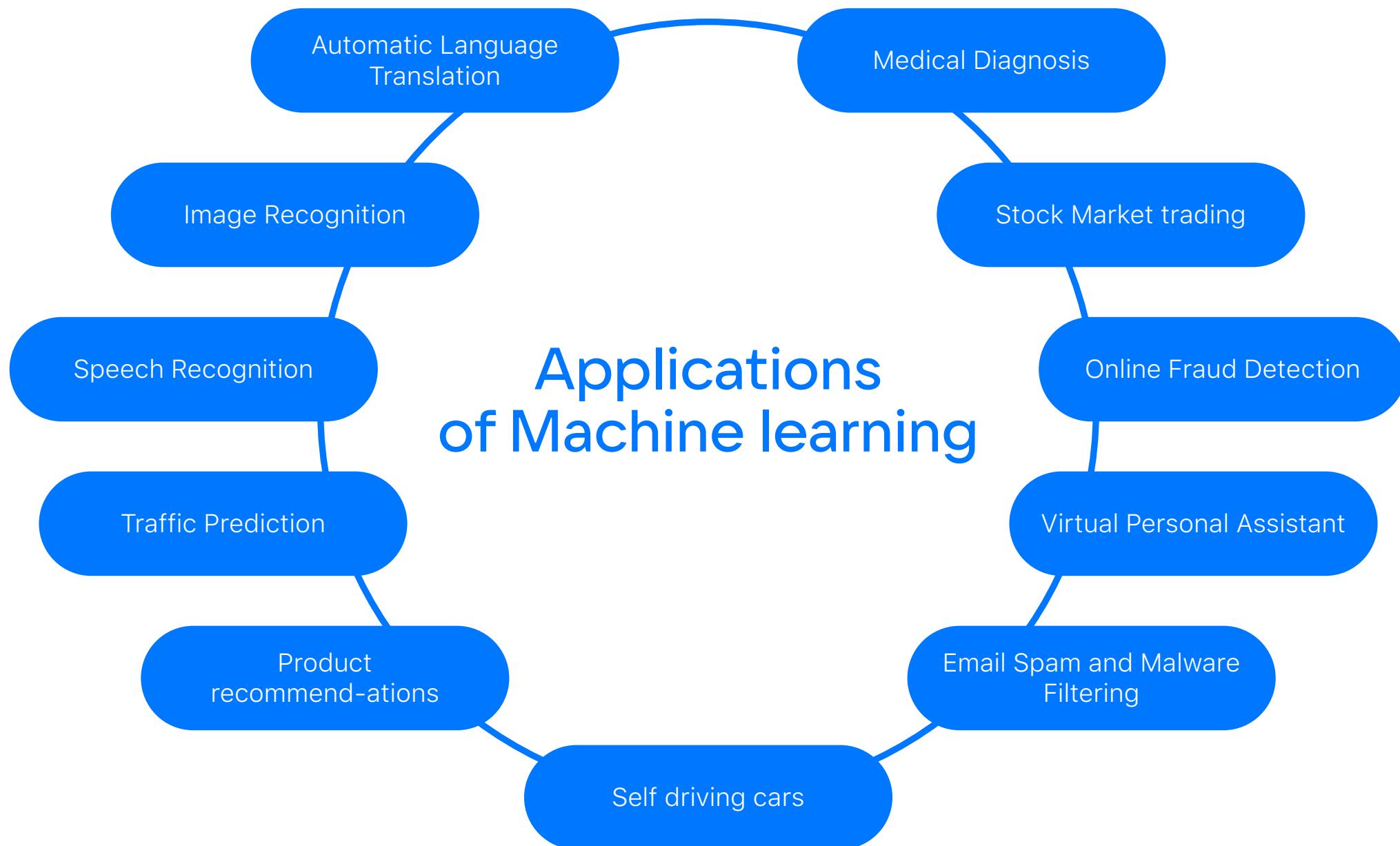
2

Обучение с подкреплением (reinforcement learning)

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

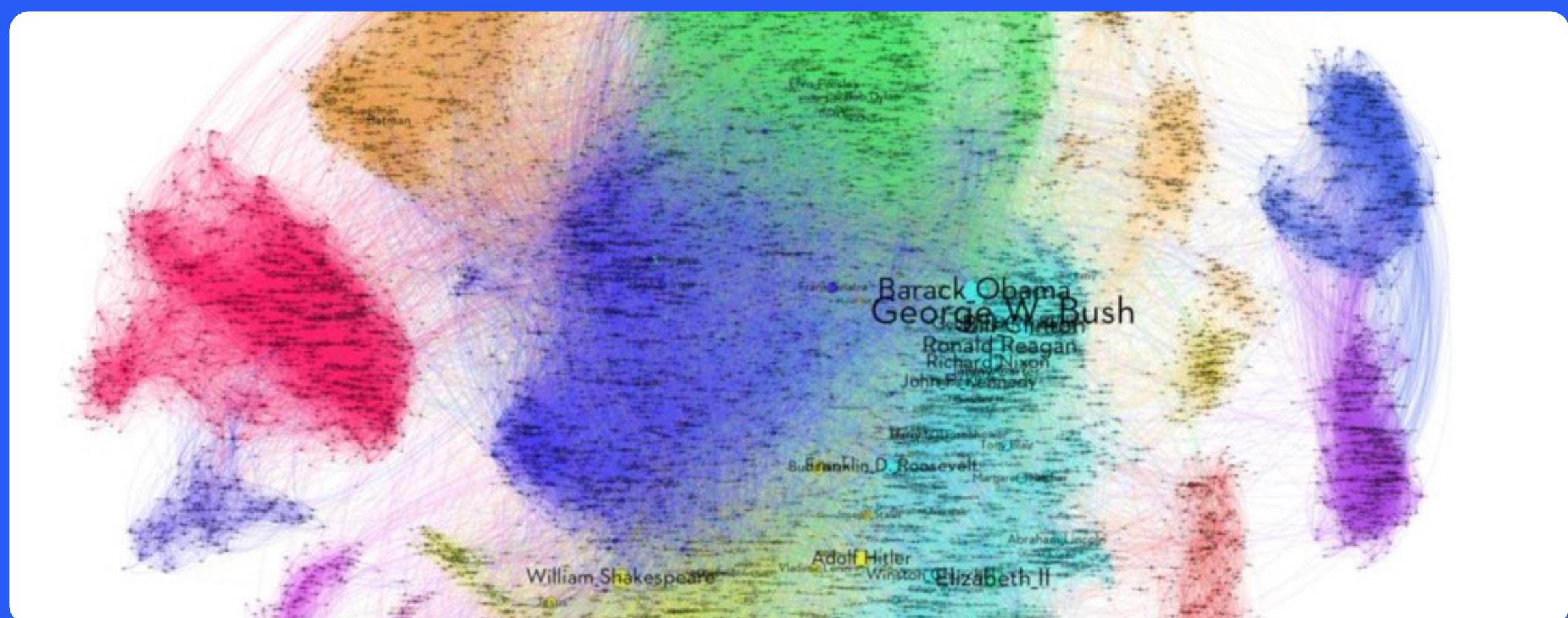
4

Applications of Machine learning



Пример обучения без учителя

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



Пример обучения без учителя

Поиск музыки одинакового жанра



Пример обучения без учителя

Рекомендательные системы



КиноПоиск

▶ Онлайн-кинотеатр Активировать промокод

Фильмы, сериалы, персоны

Рекомендации

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

Фильмы Сериалы

С высоким рейтингом

Российские Зарубежные

Скрыть просмотренные

Жанры

Все жанры

Страны

Все страны

Годы

Все годы

Онлайн Все

84 фильма 100 фильмов

По порядку

1 Игра престолов Game of Thrones, 2011–2019 9.0 535 974

США, Великобритания • фэнтези, драма

По подписке Амедиатека+КиноПоиск

2 Джентльмены Gentlemen, The, 2019 8.5 315 323

Великобритания, США • боевик, комедия

Смотреть В подписке

3 Зеленая книга Green Book, 2018 8.3 336 109

США, Китай • комедия, драма

Смотреть В подписке

Рекомендации

Рекомендации

Пример обучения без учителя

Рекомендательные системы

Всё ли нормально?



КиноПоиск ≡

▶ Онлайн-кинотеатр Активировать промокод

Фильмы, сериалы, персоны

🔍

Рекомендации

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

Рекомендации

Фильмы Сериалы

С высоким рейтингом

Российские Зарубежные

Скрыть просмотренные

Жанры

Все жанры

Страны

Все страны

Годы

Все годы

Онлайн Все

84 фильма 100 фильмов

По порядку

1 **Игра престолов**
Game of Thrones, 2011–2019
США, Великобритания • фэнтези, драма
9.0 535 974

2 **Джентльмены**
Gentlemen, The, 2019
Великобритания, США • боевик, комедия
8.5 315 323

3 **Зеленая книга**
Green Book, 2018
США, Китай • комедия, драма
8.3 336 109

Буду смотреть ⚡ ☆ ⚡ :

Буду смотреть ⚡ ☆ ⚡ :

Буду смотреть ⚡ ☆ ⚡ :

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)

- $F_j = \{\text{true}, \text{false}\}$ – 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)

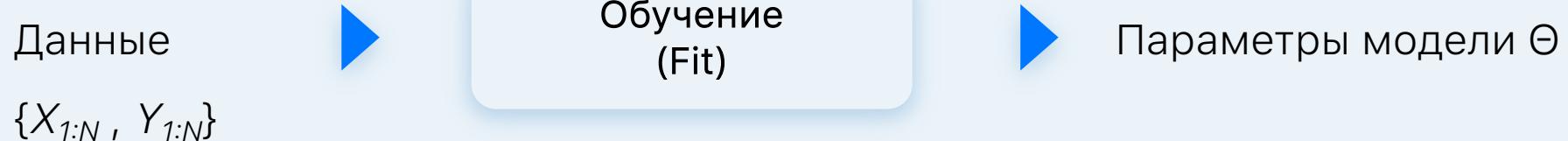
- $F_j = \mathbb{R}$ или $F_j = \mathbb{RM}$ (ответом является действительное число или числовой вектор)

Задача ранжирования (learning to rank)

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

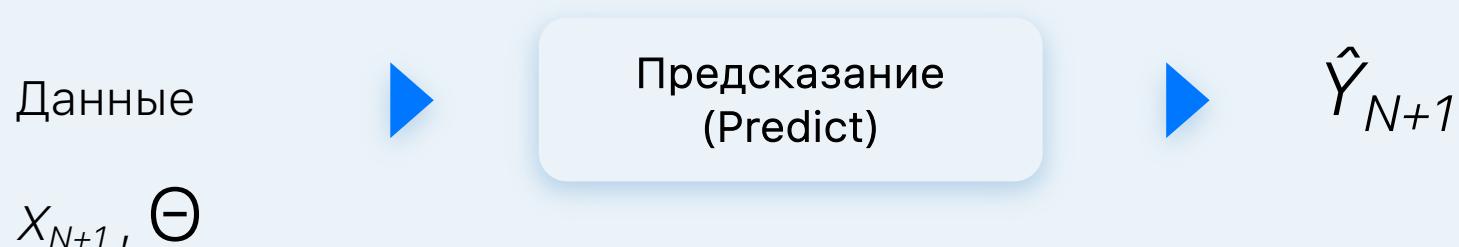
Обучение с учителем (обучения по прецедентам)

Этап обучения (train)



Необходимо учитывать представительность выборки

Этап применения (test)



Обучение с учителем (обучения по прецедентам)

Пример задачи классификации

Болеет ли человек коронавирусом?
(бинарная классификация)



Обучение с учителем (обучения по прецедентам)

Пример задачи классификации

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



Обучение с учителем (обучения по прецедентам)

Пример задачи классификации

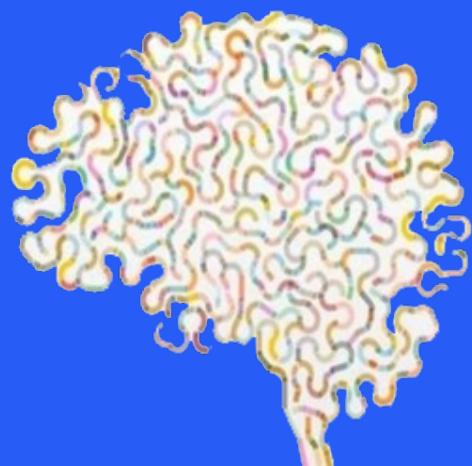
Что представлено на изображении? (классификация на M классов, которые могут пересекаться)



Обучение с учителем (обучения по прецедентам)

Пример задачи классификации

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

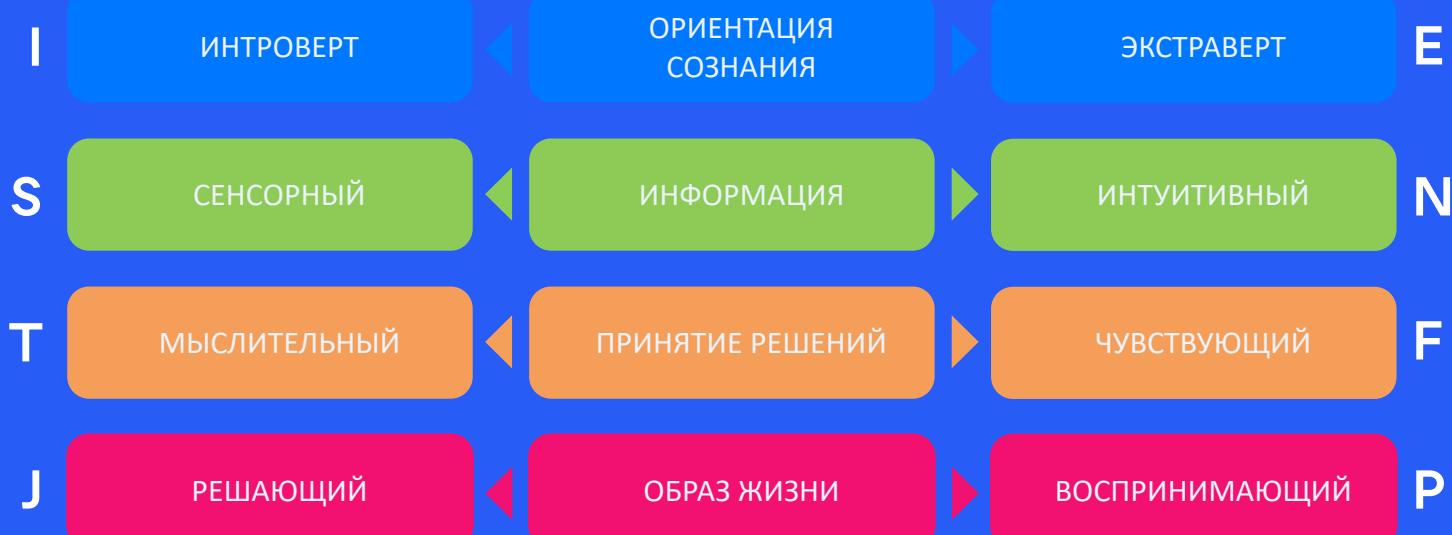


Обучение с учителем (обучения по прецедентам)

Пример задачи классификации

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

ОПРОСНИК КЕЙРСИ (ТИПОЛОГИЯ МАЙЕРС-БРИГГС)



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



```
РЕЗЮМЕ IT-ШНИКА
Objective {
    /*...*/
}

#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;
    Resposibility: 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



ер: задачи ранжирования

Картинки Новости Видео Покупки Ещё Настройки Инструменты

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

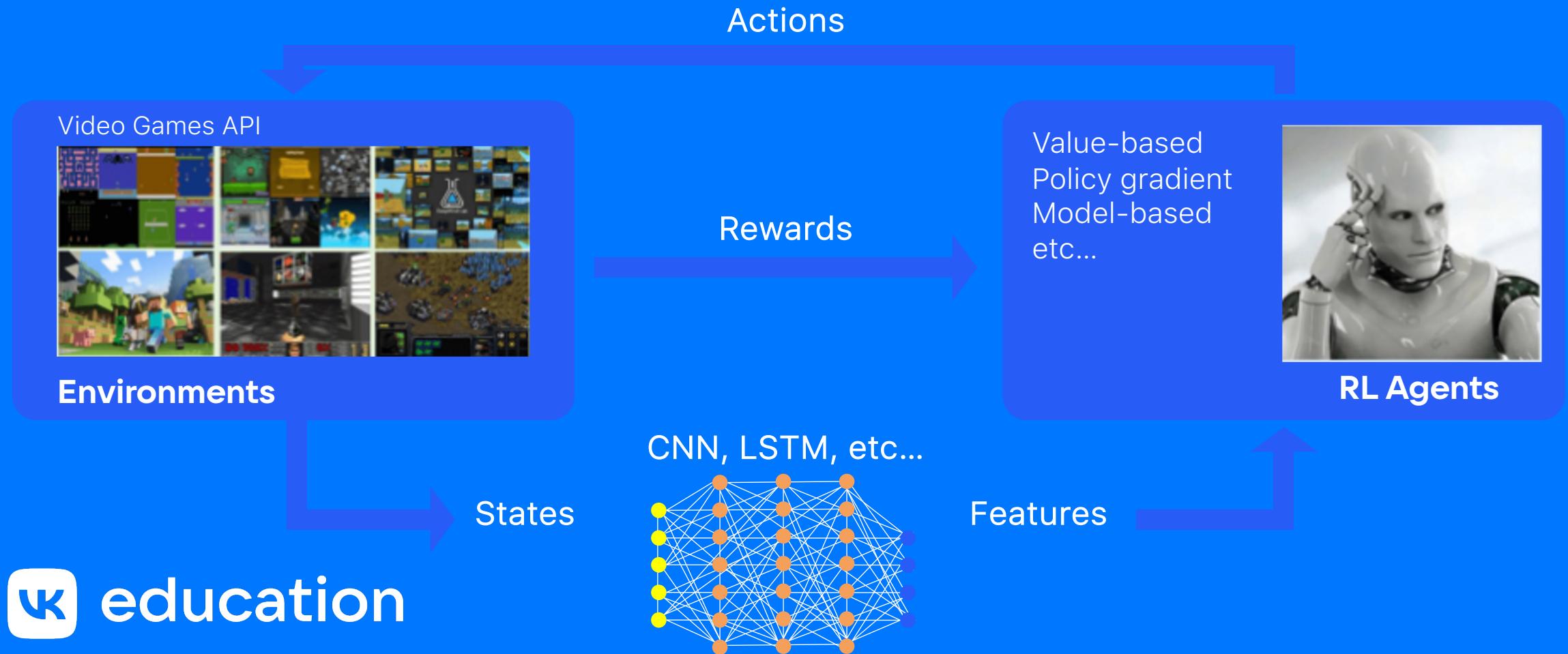
Оценки качества. ◦ Подходы к решению задачи. – поточечный ... 9. Пример вычисления nDCG ... Сведем задачу ранжирования к задаче предсказания.

www.hse.ru › data › 2012/06/20 › Алгоритмы ранжирования и их п... PDF

Алгоритмы ранжирования и их применение в задачах ...

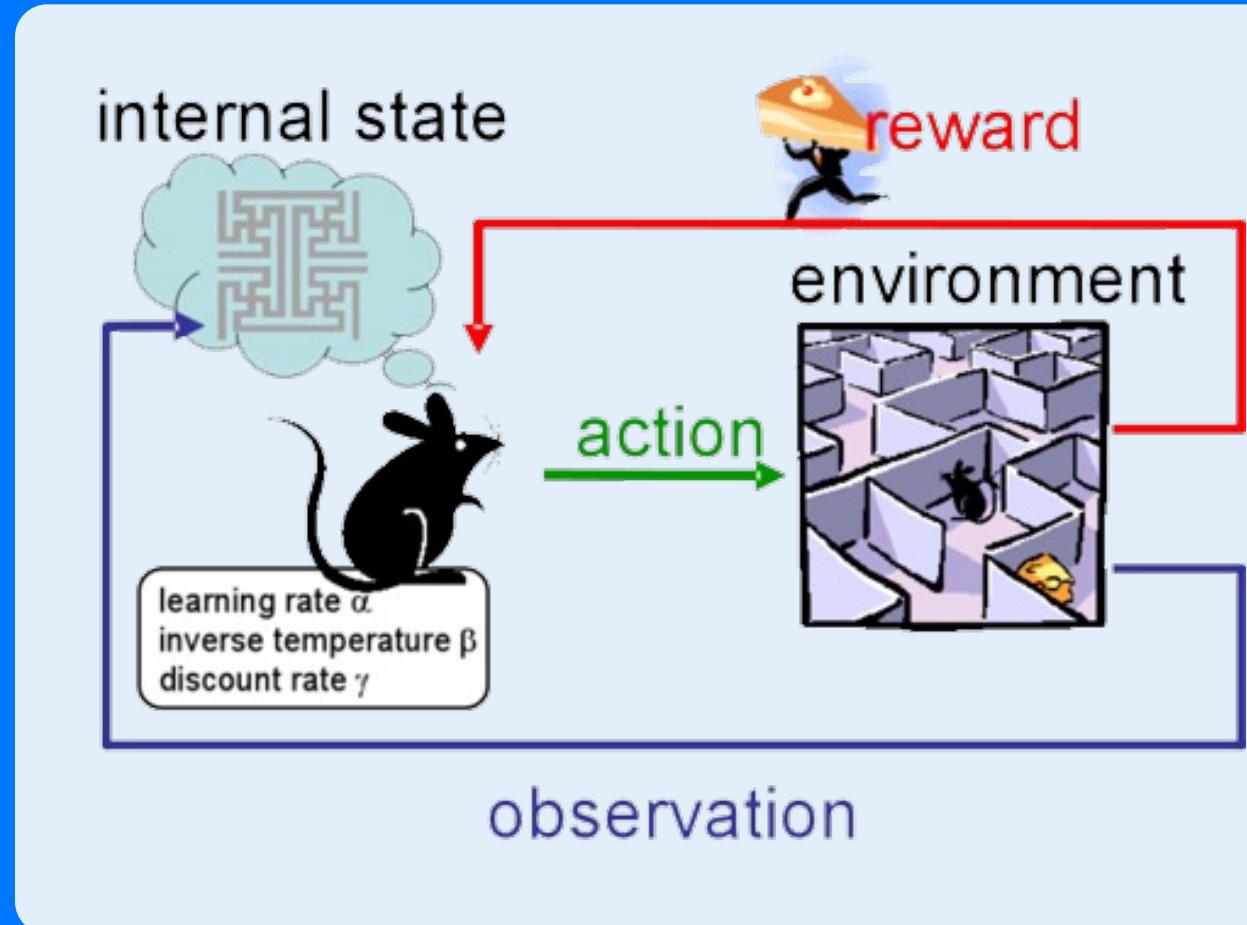
20 июн. 2012 г. - системах. Задачи работы: 1) Выявление существенных факторов, влияющих на ранжирование. 2) Кластеризация схожих запросов.

Обучение с подкреплением (reinforcement learning)



Обучение с подкреплением (reinforcement learning)

Пример: игры atari



education

Обучение с подкреплением (reinforcement learning)

Пример: создание чатбота



Инструменты

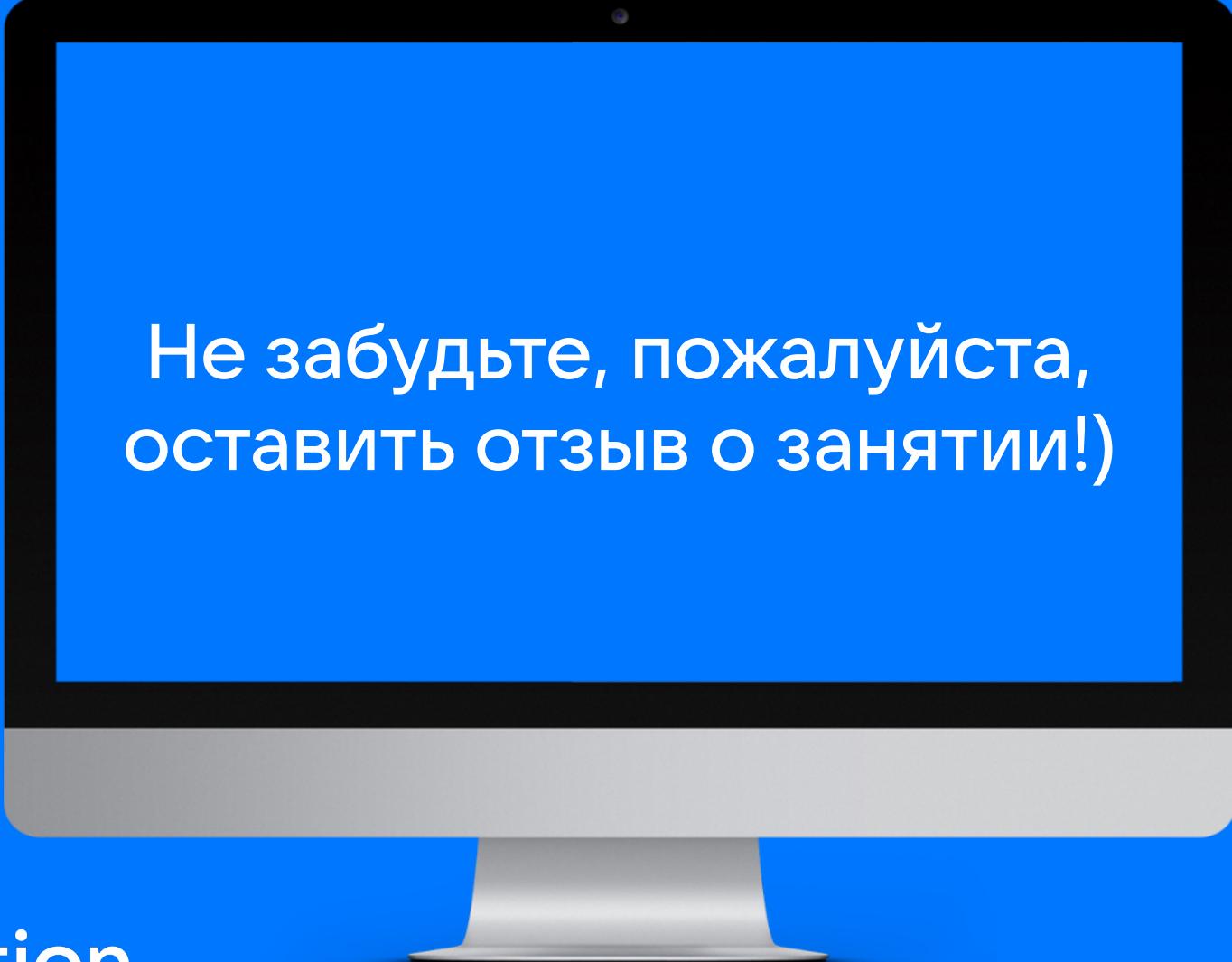


Инструменты



Python инструменты





Не забудьте, пожалуйста,
оставить отзыв о занятии!)

Введение в анализ данных

Практическая часть