

# Основные понятия машинного обучения и анализа данных

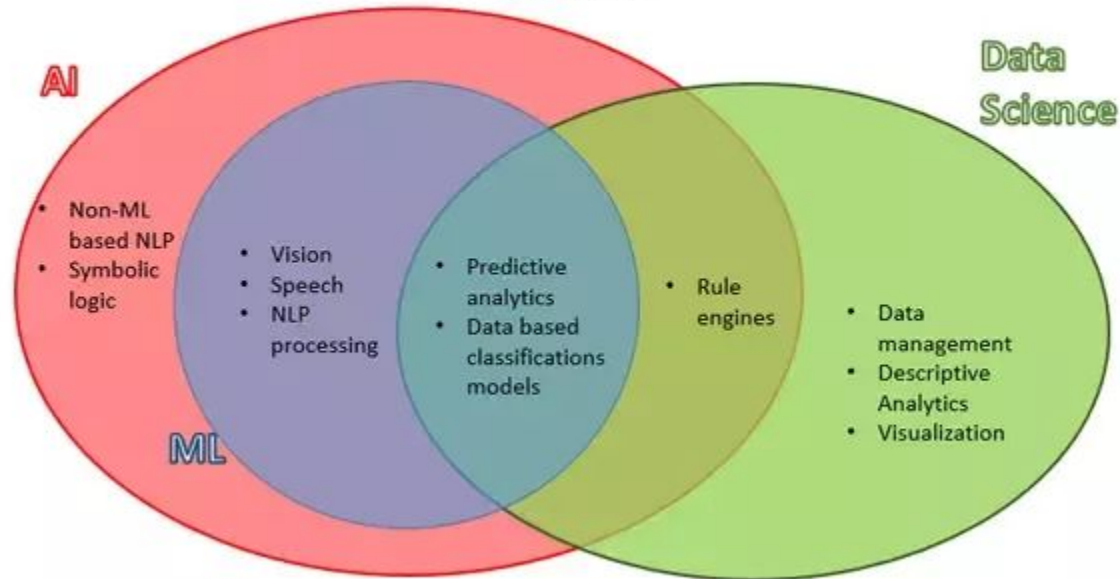
к.ф.-м.н., заместитель руководителя по научной работе, доцент

ДАДиМО

Корчагин С.А.

[SAKorchagin@fa.ru](mailto:SAKorchagin@fa.ru)

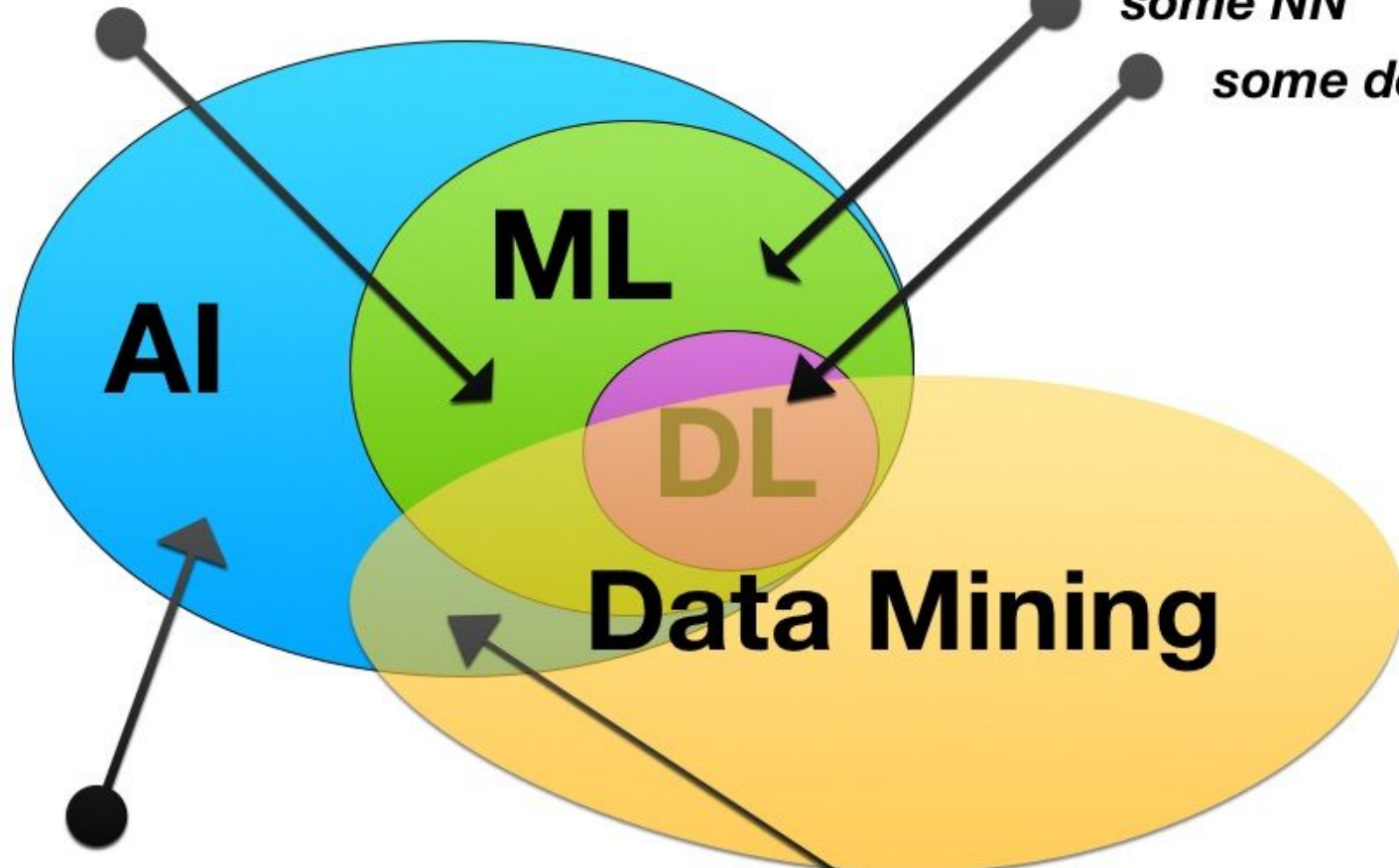
## Computational Cognitive Sciences



*e.g. genetic algorithms*

*some NN*

*some deep NN*



**AI**

**ML**

**DL**

**Data Mining**

*e.g. rule based engine*

*e.g. classification algorithms*

Говорят, что компьютерная программа обучается на основе опыта  $E$  по отношению к некоторому классу задач  $T$  и меры качества  $R$ , если качество решения задач из  $T$ , измеренное на основе  $R$ , улучшается с приобретением опыта  $E$ .

# Области применения ML

Распознавание речи.

Компьютерное зрение.

Компьютерная лингвистика и обработка естественных языков.

Медицинская диагностика.

Техническая диагностика.

Рубрикация текстов.

Интеллектуальные игры.

# Главные вопросы ML

Какое количество и какой информации необходимо для обучения?

Какие данные лучше выбирать для обучения и почему?

Какой алгоритм решает поставленную задачу наилучшим образом?

Как свести какую-либо задачу обучения к аппроксимации или оптимизации некоторой функции?

# ОСНОВНЫЕ ПОНЯТИЯ

Признак (feature)

Объект (object)

Чистые данные (tidy data)

Набор данных (dataset)

Модель

Шкала

# Определение чистых данных (tidy data)

Каждая переменная соответствует колонке

Каждое измерение соответствует строке

Каждая таблица\файл содержит данные об одном виде наблюдений\экспериментов



# Обзор данных (data exploration)

Отсутствующие данные

Значения вне разумного диапазона

Ошибки в единицах измерения (шкалах)

Ошибки в подписях переменных (колонок)

Ошибки в классах переменных

# Предварительная обработка данных

Создание новых переменных

Слияние наборов данных

Трансформация переменных

Удаление несогласованных данных

# Этапы анализа данных

Определить вопрос

Определение идеального набора данных

Определение доступного набора данных

Получение данных

Очистка данных

Исследовательский анализ данных

Статистическое моделирование

Интерпретация результатов

Проверка результатов

Описание результатов

Создание воспроизводимого кода

# Этапы процесса машинного обучения

Получение данных

Трансформация данных

Очистка данных

Визуализация данных

EDA

Выбор модели

Обучение модели

Верификация результата

# Основные типы шкал

Бинарные (Пол, наличие боли в спине, в сознании ли пациент).

Номинальные (Тип боли: колющая, режущая, ноющая).

Порядковые (Общее состояние больного: удовлетворительное, средней тяжести, тяжелое, крайне тяжелое).

Количественные (Температура тела, пульс, артериальное давление).

# Основные форматы хранения наборов данных

CSV

XML

JSON

XLSX

DB

# Главные задачи машинного обучения

Обучение с учителем (supervised learning)

- Регрессия

- Классификация

Обучение без учителя (unsupervised learning)

- Понижение размерности

- Обнаружение аномалий

- Кластеризация

Рекомендательные системы

Обучение с подкреплением (reinforcement learning)

Основой машинного обучения является оптимизация некоторой функции ошибки





# scikit-learn algorithm cheat-sheet

