Титульный лист материалов по дисциплине

(заполняется по каждому виду учебного материала)

ДИСЦИПЛИНА	**Технологии извлечения знаний из больших данных** (полное наименование дисциплины без сокращений)
ИНСТИТУТ	ИКБ
КАФЕДРА	**КБ-4 «Интеллектуальные системы информационной безопасности»** полное наименование кафедры)
ВИД УЧЕБНОГО МАТЕРИАЛА	**Лекция** (в соответствии с пп.1-11)
ПРЕПОДАВАТЕЛЬ	**Никонов В.В.** (фамилия, имя, отчество)
СЕМЕСТР	**3 семестр 2023/2024 уч. года** (указать семестр обучения, учебный год)

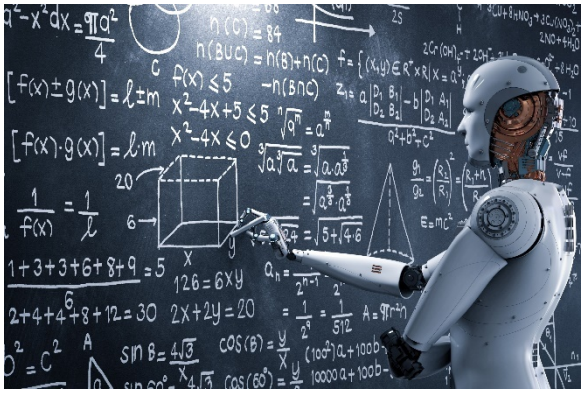
Введение

Мы начинаем курс, посвященный применению интеллектуальных методов извлечения знаний. Извлечение знаний основано на теории искусственного интеллекта. Искусственный интеллект — это сложное понятие, для которого не существует однозначного определения. Понятие искусственного интеллекта (для удобства сокращают как «ИИ») используется специалистами в различных областях: писателями, журналистами, в бизнесе и науке; и разные специалисты вкладывают свой смысл в это понятие.

В самом широком смысле искусственным интеллектом называют способность компьютера решать те же интеллектуальные задачи, которые способен решать человек.

Указанное понятие можно конкретизировать на разных уровнях:

- машина, способная воспринимать и понимать мир через сенсоры (например, анализ изображений и звука);
- способная придумывать и создавать новые объекты (например, изображения, видео и тексты);
- способная решать интеллектуальные задачи (например, игра шахматы или го);
- или способная переключаться между задачами и творчески решать сложные интеллектуальные задачи.



На сегодня существует только узкоспециализированный искусственный интеллект: машина, способная решать одну данную интеллектуальную задачу (например, распознавание лиц, игра в шахматы или

машинный перевод). При этом неясно, возможно ли в принципе с использованием существующих технологий создать общий искусственный интеллект, то есть машину, которая сможет решать различные сложные интеллектуальные задачи.

Возьмем персонального помощника: общий ИИ подразумевает, что машина сможет заменить помощника целиком (будет выполнять различного рода поручения, планировать расписание, отвечать на звонки и так далее, а также сможет самостоятельно обучаться новым задачам). Существующие сегодня технологии гораздо проще: машина может записывать текст по речевому вводу, предлагать короткие ответы на входящие письма вида «Принято в работу. С уважением, (имя)», или устанавливать напоминания по текстовому или речевому описанию, причем за каждую перечисленную функцию отвечает отдельный алгоритм.

С другой стороны, создание алгоритмов выполнения даже этих задач, кажущихся несложными на контрасте с функционалом полноценного персонального помощника, исторически потребовало немало времени.

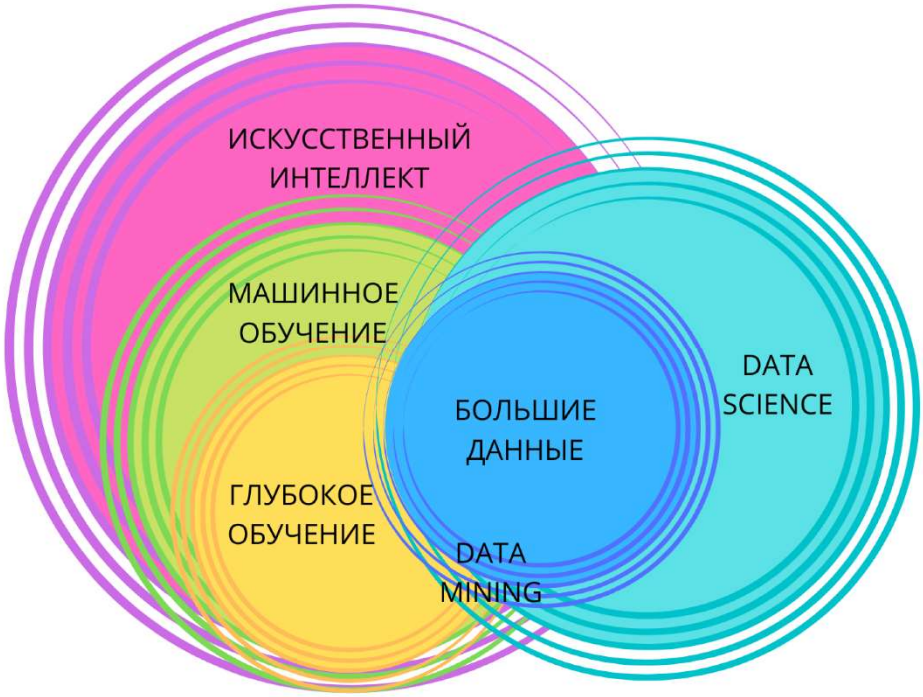
1. Области искусственного интеллекта

Искусственный интеллект включает в себя множество областей математики и информационных технологий, а также биологии, физики и других наук. Мы описали только самые известные и значимые события и подходы, разумеется, идей создания искусственного интеллекта было гораздо больше. Например, существует подход на основе эволюционных алгоритмов: он заключается в том, чтобы попытаться имитировать «эволюцию» с

помощью случайных «мутаций» программы; сегодня такие подходы используются совместно с современными системами искусственного интеллекта, например с нейронными сетями.

Помимо понимания искусственного интеллекта как способности компьютера решать интеллектуальные задачи подобно человеку, существует понимание ИИ как создания компьютера, имитирующего человеческий мозг. Однако в науке пока нет полного понимания, как работает мозг, поэтому способов его искусственного повторения тоже не существует.

На сегодняшний день технологии искусственного интеллекта и обработки больших объемов данных активно используются в бизнесе, конкретным примерам посвящен следующий блок курса. При этом выделяют несколько смежным областей, отвечающих за разработку этих технологий.



Машинное обучение

Класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение в процессе применения решений множества сходных задач.

Глубинное обучение

Иногда называют «глубокое обучение» (от англ. Deeplearning). Подобласть машинного обучения, где в качестве алгоритмов используются нейронные сети.

DataScience

Это концепция объединения статистики, анализа данных, машинного обучения и связанных с ними методов для понимания и анализа реальных явлений.

DataMining

Широкое понятие, означающее извлечение знаний из данных.

Большие данные

Это набор подходов и методов, разработанных для анализа данных огромных размеров.

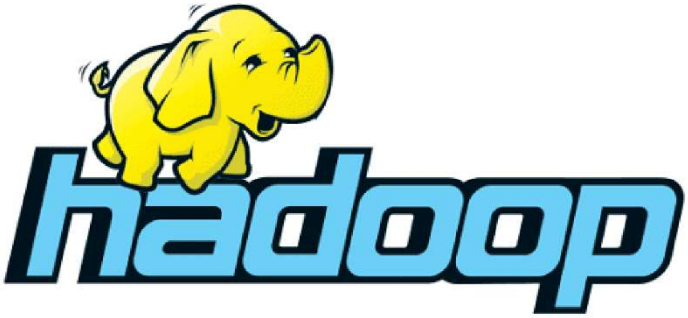
Стоит отметить, что не всякая работа с данными относится к искусственному интеллекту (например, аналитик, делающий вывод по графикам, не относится к искусственному интеллекту), и не все алгоритмы искусственного интеллекта разрабатываются с использованием данных (например, экспертные системы, упомянутые выше). Тем не менее, множество современных интеллектуальных систем основано именно на обучении по данным: машинный перевод, распознавание изображений и речи, прогнозирование поведения клиентов и др. Во время обучения по данным алгоритм «изучает» большое количество реальных случаев (например, поведения клиентов или переводов текстов) и благодаря этому делает качественные предсказания для новых случаев.

Сегодня технологии искусственного интеллекта во многом основываются на обучении по большим объемам данных, однако помимо этого также включают сложные вычисления, экспертные системы и другие алгоритмы.

2. Технологии работы с большими данными

Современные технологии практически целиком основываются на работе с данными, и, кроме того, использование данных и аналитика данных сами по себе являются мощным драйвером развития в современном мире. При этом речь идет об огромных массивах данных, обработать которые на персональном компьютере не предоставляется возможным. Об инструментах хранения и обработки больших данных мы говорили в соответствующем курсе, здесь кратко напомним перечень.

Для хранения и обработки больших данных создают специальные дата-центры или арендуют мощности на облачных сервисах, таких как AmasonWebServices или MicrosoftAzure.



Вместо последовательной обработки данных о миллиардах клиентов используют распределенное хранение данных и параллельную их обработку, например в Hadoop или Spark. Отдельный модуль SparkML позволяет строить модели машинного обучения на сверхбольших объемах данных.

Для построения моделей машинного обучения наиболее популярен язык программирования Python и его библиотеки Scikit-learn, LightGBM, CatBoost и другие. Для обучения нейронных сетей используются отдельные библиотеки: PyTorch и TensorFlow. Также существуют графические интерфейсы, напримерRapidMiner, но они предоставляют более ограниченный функционал по сравнению с Python и поэтому не так популярны.