



Конференция «Программирование и вычислительная математика»,
посвящённая 100-летию со дня рождения Николая Павловича Трифонова

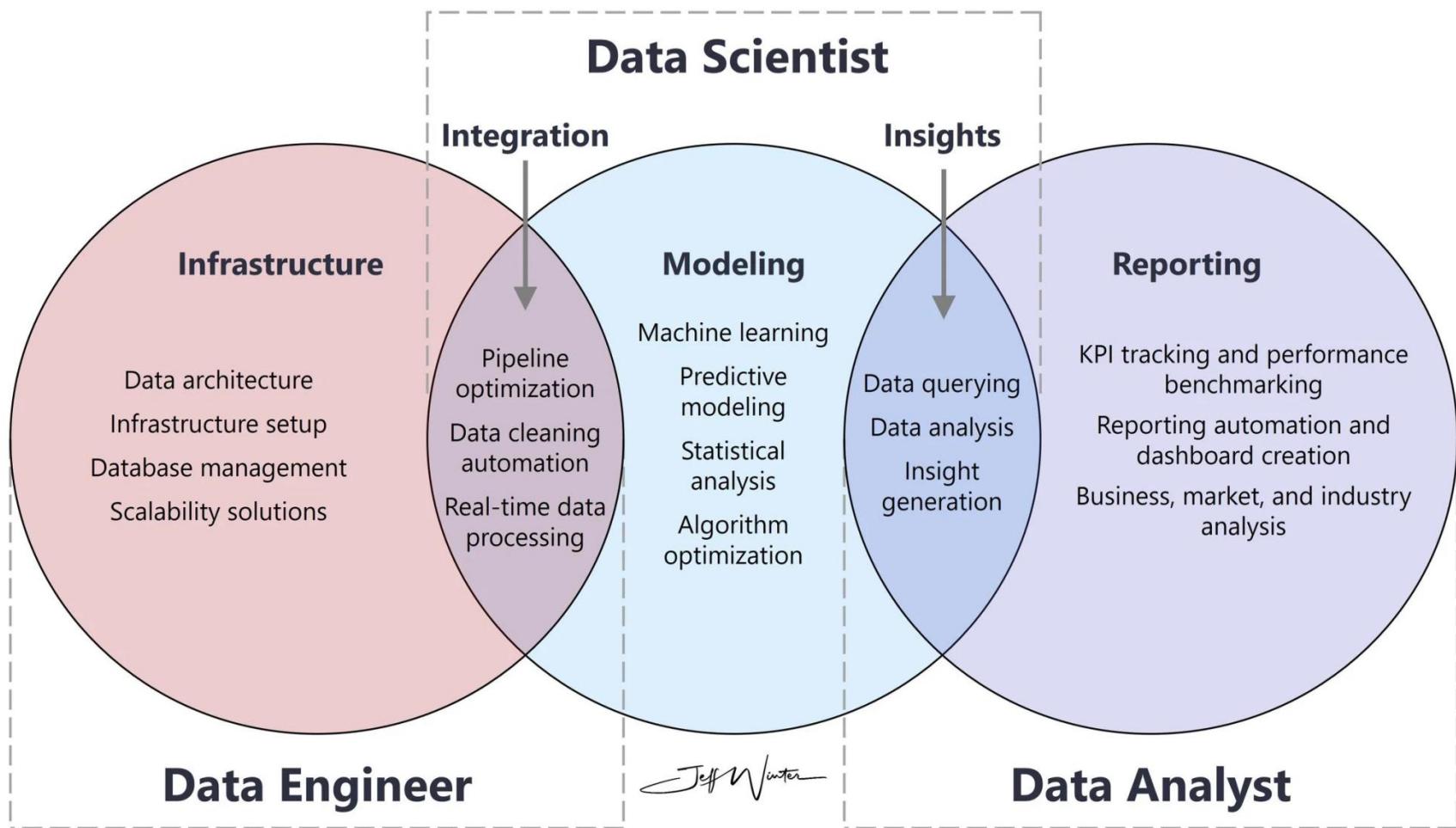


СОВЕРШЕНСТВОВАНИЕ МЕТОДИКИ ПОДГОТОВКИ СТУДЕНТОВ МАТЕМАТИЧЕСКИХ СПЕЦИАЛЬНОСТЕЙ ДЛЯ РЕШЕНИЯ ЗАДАЧ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И НАУКИ О ДАННЫХ

*кафедра теоретической информатики
механико-математического факультета МГУ
к.ф.-м.н., доц. Главацкий Сергей Тимофеевич,
ст.преп. Бурыкин Илья Геннадиевич*

- В настоящее время подготовка специалистов по искусственному интеллекту (ИИ), методам машинного обучения, интеллектуальному анализу данных и работе с большими данными становится все более востребованной и распространенной.
 - **Однако существующее обучение зачастую имеет скорее прикладную, а не фундаментальную направленность.**
 - **В то же время становится понятным, что успехи, достигнутые в совершенствовании методов искусственного интеллекта, в частности в машинном обучении, еще не являются основой для построения полноценной научной теории в исследуемой предметной области.**
 - Мы развиваем свою образовательную траекторию, ориентированную **на подготовку ученых по данным.**

Статус ученого по данным



Adaptation of the data science diagram created by Kevin Schmidt in Towards Data Science, 2015

Статус ученого по данным

- **Подготовка специалистов по ИИ должна быть выделена в самостоятельную научно-образовательную область, в фундаменте которой находится, прежде всего, наука о данных** наряду с примыкающими областями знаний.
- Проблемы обучения студентов «Науке о данных» и подготовки специалистов по ИИ **имеют свою специфику для классических университетов**, готовящих специалистов в рамках программ академического высшего образования.

Анализ данных: от постановки задачи до ее решения и осмысления результатов

- Поскольку в основе интеллектуального анализа данных (в частности, в методах и средствах ИИ), как и в других компьютерных реализациях, лежит известная **триада академика А.А. Самарского «модель – алгоритм – программа»**, то любой пользователь методов ИИ (даже далекий от математики и информатики) должен понимать, что **в основе исследования изначально лежит математическая модель.**
- А выбор модели и, далее, метода обработки данных, построения/адаптации эффективного алгоритма, программной реализации – это удел специалиста, знающего фундаментальные математические основы методов решения поставленных задач.
- При этом на его выбор, безусловно, влияет мнение эксперта в данной предметной области. Именно содружество специалистов из различных областей знания является ключом к построению верных и эффективных решений в сфере ИИ.

Анализ данных: математическая модель

IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, VOL. 1, NO. 1, APRIL 1997

No Free Lunch Theorems for Optimization

David H. Wolpert and William G. Macready

Теорема (No free Lunch Theorem):

Пусть $P(d_m^y | f, m, a)$ — условная вероятность получения частного решения d_m после m итераций работы алгоритма a при целевой функции f . Для любой пары алгоритмов a_1 и a_2 имеет место равенство:

$$\sum_f P(d_m^y | f, m, a_1) = \sum_f P(d_m^y | f, m, a_2)$$

- Теоремы NFL означают, что если алгоритм в среднем показывает особенно хорошие результаты для одного класса задач, то в остальных задачах он должен показывать худшие результаты.

Анализ данных: математическая модель

- <https://konfuzio.com/en/sate-of-the-art/>
- Что такое SOTA? Сокращённо «State-of-the-Art» означает **наиболее эффективную модель или алгоритм для конкретной задачи или области исследований** в области машинного обучения.
- SOTA представляет собой самый передовой и современный подход, обеспечивающий высочайшую точность или исключительную функциональность.

Анализ данных: этапы процесса решения задач

- **Восприятие и анализ постановки задачи:** тщательное изучение и интерпретация условий и требований к методам решения.
 - Этот этап включает в себя понимание контекста задачи, определение ее целей, выявление ключевых переменных и ограничений;
- **Построение модели явления:** формализация исходных данных и ограничений для дальнейшей работы.
 - На этом этапе данные преобразуются в формат, пригодный для машинной обработки, выбираются подходящие математические и алгоритмические модели, которые смогут отразить сущность исследуемого явления;

Анализ данных: этапы процесса решения задач

- **Поиск способа решения задачи:** выбор оптимального алгоритмического подхода или комбинации методов.
 - Здесь происходит подбор соответствующих алгоритмов, методов машинного обучения, статистических техник, которые наилучшим образом соответствуют поставленной задаче и имеющимся данным;
- **Реализация выбранного метода решения задачи:** применение выбранного алгоритма.
 - Этот этап включает в себя написание кода, настройку параметров моделей, выполнение вычислений.

Анализ данных: этапы процесса решения задач

- **Проверка выбранного метода решения задачи:** верификация корректности полученного результата и оценка его качества.
 - Здесь проводится тестирование модели, анализ метрик производительности, проверка соответствия результатов исходным условиям и здравому смыслу;
- **Формулировка ответа:** представление результатов решения задачи в понятной и доступной форме.
 - Это может быть отчет, презентация, визуализация данных, рекомендация или готовый продукт;

Анализ данных: этапы процесса решения задач

- **Учебно-познавательный анализ поставленной задачи и ее решения:** рефлексия над процессом решения, извлечение обучающих выводов и обобщение полученного опыта.
 - Этот этап включает в себя анализ допущенных ошибок, выявление лучших алгоритмов, осмысление примененных методов.

Цикл спецкурсов «Аналитика больших данных для математиков»: содержание и структура

- На кафедре теоретической информатики мех-мата МГУ более десяти лет назад был разработан цикл специальных курсов и практикумов под общим наименованием «Аналитика больших данных для математиков» («Data Science and Data Mining for Mathematicians»).

 **Современные
информационные
технологии
и ИТ-образование**
Международный научный журнал

[О журнале](#) • [Архив](#) [Текущий выпуск](#)

 29 мая 2017 г. журнал включен в Перечень ведущих рецензируемых научных журналов и изданий, рекомендованных ВАК РФ для публикации основных результатов диссертационных исследований.

[Главная](#) / [Архив](#) / Том 12 № 3-2 (2016): Современные информационные технологии и ИТ-образование / ИТ-образование: методология, методическое обеспечение

О ЦИКЛЕ КУРСОВ «АНАЛИТИКА БОЛЬШИХ ДАННЫХ ДЛЯ МАТЕМАТИКОВ»

Сергей Тимофеевич Главацкий
Московский государственный университет имени М.В. Ломоносова

Илья Геннадиевич Бурыкин
Московский государственный университет имени М.В. Ломоносова

Аннотация

В статье излагается подход к преподаванию специализации в области науки о данных для математиков на кафедре теоретической информатики механико-математического факультета МГУ имени М.В. Ломоносова. Предложен авторский взгляд на выбор тем и курсов (как основных, так и специальных).


ISSN 2085-9473
Международный научный журнал
**Современные
информационные
технологии
и ИТ-образование**
Том 12, номер 3, часть 2, 2016

Цикл спецкурсов «Аналитика больших данных для математиков»: содержание и структура

- В рамках этого образовательного направления на базе имеющихся общих курсов, а также за счет введения новых междисциплинарных предметов, авторы осуществляют преподавание методов и алгоритмов для представления, моделирования и анализа больших наборов данных.
- Образовательная программа сочетает общие дисциплины с новыми междисциплинарными предметами.
- Среди полугодовых специальных курсов:
 - «Аналитика больших данных: основные алгоритмы»,
 - «Аналитика больших данных: дополнительные главы»,
 - «Модели данных и основы систем баз данных» и
 - «Современные технологии баз данных: от In-Memory до решений искусственного интеллекта».

Цикл спецкурсов «Аналитика больших данных для математиков»: содержание и структура

- Можно отметить следующие особенности предлагаемых спецкурсов. Они:
 - включают в себя и теоретическую, и практическую составляющие;
 - являются,
 - с одной стороны, взаимозависимыми (для полного понимания предметной области),
 - а с другой – не требуют обязательного предварительного изучения содержания остальных спецкурсов из предложенного набора;
 - отражают как уже ставшие классическими, проверенные временем модели и алгоритмы, так и актуальные современные взгляды, тенденции и понятия в области анализа данных.

Спецкурс «Аналитика больших данных: основные алгоритмы»

- Данный спецкурс знакомит студентов с современными достижениями в теории и практике обработки больших данных, выделяя методы выявления скрытых связей и закономерностей в разнородных массивах информации.
- Основные темы курса охватывают:
 - современные **методы хеширования и индексирования данных**,
 - **локально-чувствительное хеширование (LSH)** и его применение,
 - **методы оценки сходства объектов** (Euclidean distance, Hamming distance, cosine similarity и др.),
 - **кластеризацию и классификацию данных** (иерархические, k-means, BFR, CURE алгоритмы),
 - **алгоритмы поиска частых наборов элементов** (Apriori, PCY, SON),
 - **методы распараллеливания алгоритмов** обработки больших наборов данных (на примере MapReduce).
- Цель курса — сформировать у студентов глубокие знания и практические навыки, необходимые для эффективного интеллектуального анализа данных в условиях стремительно развивающихся технологий.

Спецкурс «Аналитика больших данных: основные алгоритмы»

- Особый акцент сделан на изучении новых разработок в области **приближённого поиска ближайших соседей (ANN)**, таких как **высокоэффективные структуры данных (HNSW, FAISS)** и алгоритмы оптимизации поисковых индексов, включая **метод инвертированного индексирования файлов (Inverted File Indexing, IVF)** для оптимизации работы векторных баз данных.
- Эти подходы открывают значительные перспективы для внедрения в сферы рекомендательных систем, распознавания образов и анализа семантически близких документов.

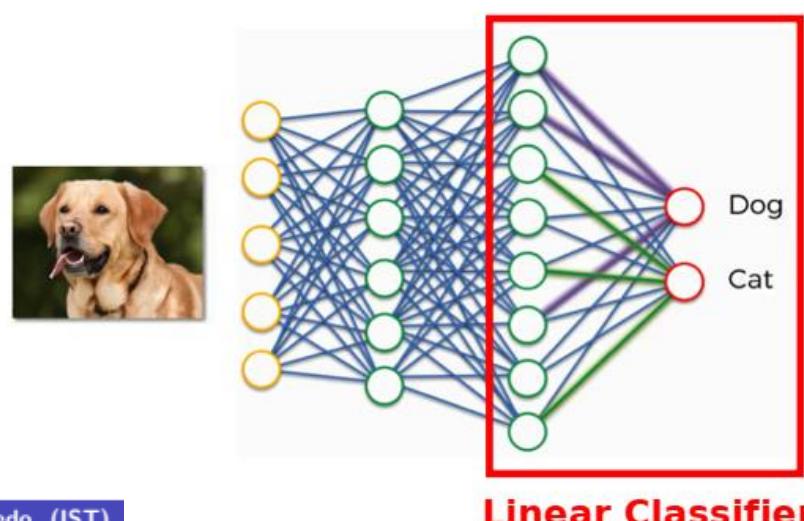
Спецкурс «Аналитика больших данных: дополнительные главы»

- Курс ориентирован на углубленное изучение ключевых методов анализа больших данных, охватывающих как классические подходы, так и новейшие технологии в области аналитики. Основное внимание уделяется таким направлениям, как:
 - **ранжирование страниц:** применение PageRank, проблемы спама и ловушек поисковых роботов;
 - **рекомендательные системы:** разработка персонифицированных рекомендаций, работа с матрицами предпочтений и контентом;
 - **классификация и кластеризация:** методы сегментации пользователей и товаров, контент-аналитика и документальная классификация;
 - **снижение размерности:** техника сингулярного разложения и метода главных компонент, защита от переобучения.
- Курс ставит своей целью научить студентов глубоко понимать технологии анализа больших данных, овладеть необходимыми практическими навыками и уметь правильно выбирать инструменты для выполнения проектов.

Спецкурс «Аналитика больших данных: дополнительные главы»

- Основываясь на нашем опыте преподавания и наблюдениях за потребностями студентов, мы пришли к выводу о необходимости дополнить программу **углубленным изучением линейных моделей**, которые являются основой машинного обучения и служат важным шагом перед изучением более сложных архитектур.

Linear Classifiers and Neural Networks



Спецкурс «Аналитика больших данных: дополнительные главы»

- Несмотря на повсеместное распространение глубоких нейронных сетей (DNN), мы убеждены, что **освоение линейных моделей является критически важным для студентов математических специальностей** по ряду причин:
 - **фундаментальность концепций:** базовые концепции машинного обучения, такие как обучение с учителем, функция потерь, оптимизация, сохраняются вне зависимости от сложности модели;
 - **теоретическая доступность:** теоретические основы линейных моделей, связанные со статистикой и оптимизацией, значительно проще для понимания студентов, что создает прочную базу для дальнейшего изучения более сложных методов;
 - **актуальность и практическое применение:** линейные модели по прежнему широко используются на практике, особенно в ситуациях, когда объемы данных ограничены, или требуется высокая интерпретируемость результатов;
 - **структурная связь с DNN:** линейные модели являются составным компонентом многих современных архитектур, включая глубокие нейронные сети, где они часто выступают в качестве выходных слоев или промежуточных преобразований.

Спецкурс «Модели данных и основы систем баз данных»

- Данный спецкурс охватывает широкий круг вопросов, начиная от базовых понятий моделирования данных и заканчивая принципами нормализации и проектированием структур баз данных. Изучаемые темы включают:
 - освоение фундаментальных концепций **моделей данных**, включая иерархическую, сетевую и реляционную;
 - понимание основ функционирования основных компонентов **системы управления базами данных** (СУБД);
 - изучение принципов построения запросов и выполнения операций над данными средствами **языка SQL**;
 - овладение методами анализа **функциональной зависимости** атрибутов и алгоритмами **нормализации схем отношений**;
 - развитие компетенций в **проектировании эффективных структур данных**, обеспечивающих целостность и эффективность обработки информации.
- Цель курса состоит в формировании глубоких теоретических знаний и практических навыков студентов в области проектирования и эксплуатации баз данных, основанных на современных моделях данных и системах управления ими.

Спецкурс «Модели данных и основы систем баз данных»

- Одним из важных аспектов курса является активное использование СУБД PostgreSQL, которая:
 - играет ведущую роль в поддержке технологического суверенитета России;
 - является идеальным инструментом для демонстрации современных реляционных моделей данных и возможностей языка SQL, позволяя студентам эффективно освоить практические навыки разработки и администрирования баз данных.

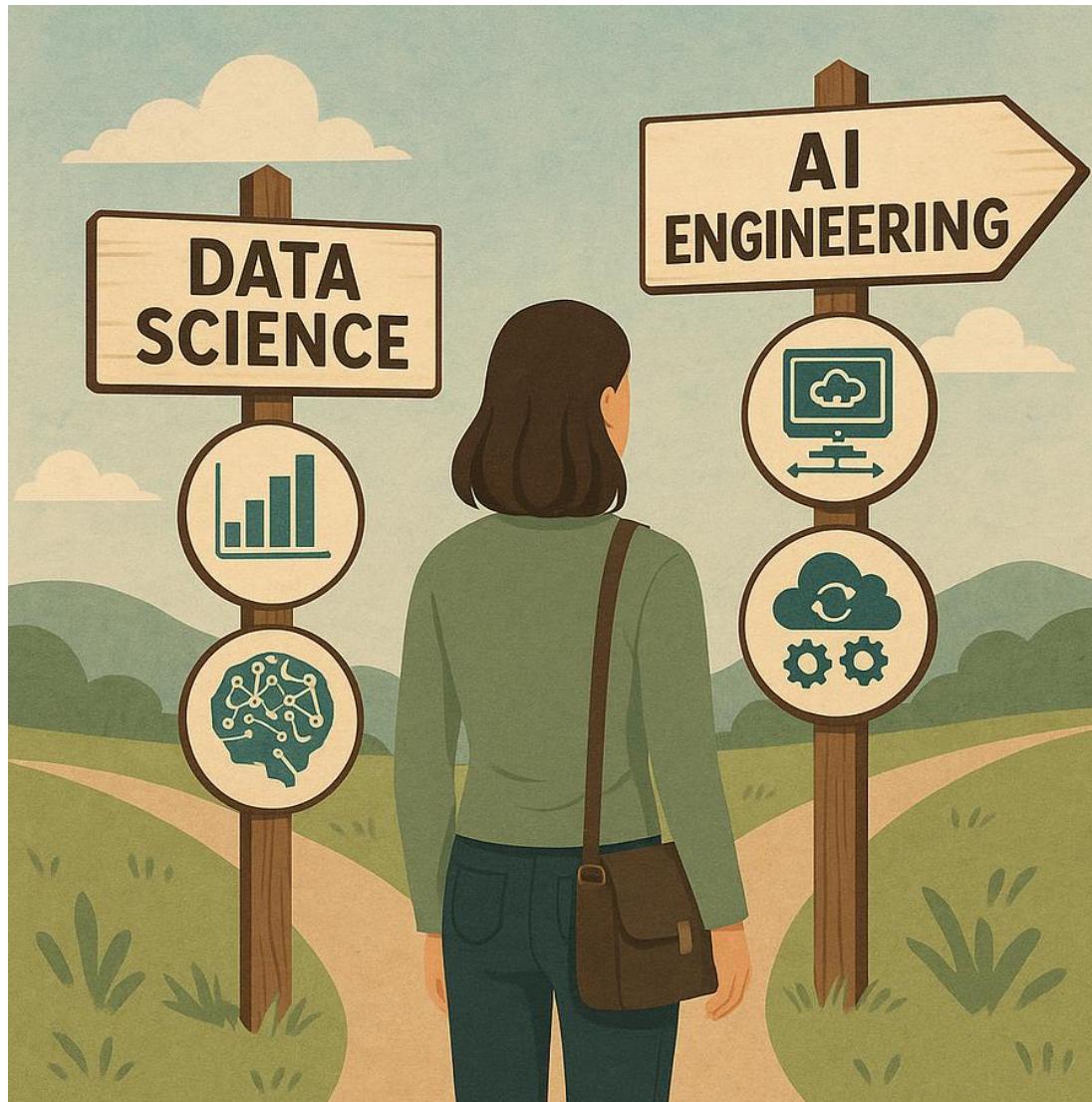
Спецкурс «Современные технологии баз данных: от In-Memory до решений искусственного интеллекта»

- Программа курса посвящена исследованию архитектуры и методов реализации **высокопроизводительных баз данных, включая In-Memory, NoSQL, NewSQL и базы данных распределённого реестра (блокчейн).**
- Основное внимание уделяется решению актуальных вопросов обработки данных на корпоративном уровне и новым технологиям управления информацией.
- Цель курса – подготовить высококвалифицированных специалистов:
 - обладающих глубокими знаниями и навыками проектирования и эксплуатации современных корпоративных приложений и баз данных;
 - способных эффективно решать реальные бизнес-задачи;
 - востребованными в условиях цифровой экономики.

Спецкурс «Современные технологии баз данных: от In-Memory до решений искусственного интеллекта»

- Ключевые нововведения:
 - **поддержка цифровых инициатив:** включены актуальные обновления, отражающие современные тенденции развития отрасли, такие как эффективные подходы к обработке больших объёмов данных;
 - **цифровой суверенитет:** представлены материалы о ведущей роли отечественной СУБД PostgreSQL в поддержке технологического суверенитета и снижении зависимости российских компаний от импортных решений;
 - **интеграция ИИ и облачных технологий:** рассмотрены концепции автономного управления базами данных с применением ИИ, адаптивной настройки и поддержки запросов, связанных с обработкой данных ИИ, а также специфика работы баз данных в облачной инфраструктуре;
 - **векторные базы данных:** изучаются специальные структуры данных и новые типы индексов, разработанные специально для ускорения работы приложений искусственного интеллекта, включая специализированные средства для хранения и быстрого поиска векторов, используемых в системах машинного обучения и анализа больших данных.

ИИ: кого нужно готовить?



AI

A high school dropout who got hired at OpenAI says he used ChatGPT to learn Ph.D.-level AI

By Lee Chong Ming



й рабочий день
текущее время - 1 г.
иу Area
rience the world

neer
ный рабочий день
к. 2024 г. - 1 г. 3 мес.

San Francisco Bay Area
takes words, produces images

[Показать все сведения об опыте работы \(16\) →](#)

Образование

Erik Dahlbergsgymnasiet
High School Diploma, Technical Science
2017 - 2019

straight As until dropping out of high school

Georgetown University
Summer Scholarship
2019 - 2019
Summer Scholarship

Парень отчислился из университета и научился кодить на уровне синьора с помощью ChatGPT. Теперь он [работает](#) исследователем в OpenAI.

Габриэль начинал в небольшом стартапе и осваивал разработку по ходу дела. После релиза ChatGPT он начал писать софт с чат-ботом — это помогло изучить базу программирования.

За пару лет Габриэль стал специалистом по машинному обучению: он был разработчиком в Midjourney, а сейчас занимает должность уровня PhD в команде Sora.

Кажется, с чат-ботами уже не страшно прогуливать пары.

ИИ: от исследований к реализации

- Развитие современных технологий искусственного интеллекта активно стимулируется увеличением вычислительных мощностей, таких как GPU, объемы памяти и доступ к электроэнергии. **Эти факторы обуславливают переход научных исследований из академической среды университетов в крупные корпорации, обладающие ресурсами для масштабирования моделей машинного обучения.**
- Концентрация инноваций в крупных компаниях формирует новую парадигму развития науки и технологий, смещающую традиционные модели финансирования и организации исследовательской деятельности.

ИИ: от исследований к реализации

ПРОКАЧИВАЕМ СКИЛЛЫ будущих AI-специалистов



Вчера в Москве завершилась двухнедельная Проектная школа AI360: 95 второкурсников из МФТИ, ВШЭ, ИТМО и Иннополиса провели это время в офисах Сбера и Яндекса — полное погружение в культуру бигтехов.

С 17 по 28 ноября они жили в режиме нон-стоп: “ первая неделя — в офисе Яндекса, вторая — в Сбере.

Разборы научных статей, командная проектная работа под менторством экспертов, встречи с топ-менеджерами и выступление Германа Грефа — так выглядит подготовка специалистов нового поколения.

Мы ЛОМАЕМ академический шаблон, чтобы дать им почувствовать ПУЛЬС БИГТЕХА ⚡



Вот где по-настоящему готовят к будущему!

”

ИИ: от исследований к реализации

- Вероятно, возврат к университетскому уровню возможен в среднесрочной перспективе, но это потребует значительных изменений как в инфраструктурных возможностях самих вузов, так и в общей культуре корпоративного сотрудничества. **Университеты станут центром развития ИИ тогда, когда доступность вычислительной мощности сравняется либо же изменится сама природа ИИ-исследований.**
- А пока, остается лишь разработка эффективных подходов к обучению малых моделей или попытаться переместиться в сторону фундаментальных вопросов теории ИИ.

ИИ: место Университета в настоящее время

- университет вечен, а аншлаг в сфере ИИ нет, поэтому поддержка фундамента в области ИТ находится в наших руках;
- мы готовим МОЗГИ, а не специалистов для быстро меняющегося ИИ-бизнеса (здесь нам за бигтехом не угнаться, да и не надо);
- мы анализируем ситуацию с фундаментальной точки зрения (**базис — Наука о данных**), и видим **свою задачу в участии в научных исследованиях**, а не в соревновании в сфере технологий (это оставим ИТ-инженерам!).

Заключение

- Авторы в своей педагогической деятельности стараются твердо придерживаться **принципа фундаментальности изложения материала, когда наука идет впереди практики.**
- Для нас главное – это то, что **полученные в процессе обучения знания, а не только навыки, далее могут активно применяться выпускниками университета как в фундаментальной науке, так и в широком круге приложений, работающих с большими данными и ИИ.**

Спасибо за внимание!