

УДК 552.5 + 553.982:550.4:004.02 + 004.85
doi: 10.55959/MSU0579-9406-4-2025-64-1-88-96

ТЕКУЩЕЕ СОСТОЯНИЕ ПРИМЕНЕНИЯ МЕТОДОВ НАУКИ О ДАННЫХ В ГЕОХИМИИ НЕФТИ И ГАЗА

Глеб Антонович Шевченко^{1✉}, Мария Александровна Большакова²

¹ Московский государственный университет имени М.В. Ломоносова, Москва, Россия; gleb.a.shevchenko@gmail.com ✉

² Московский государственный университет имени М.В. Ломоносова, Москва, Россия; m.bolshakova@oilmsu.ru,
<https://orcid.org/0000-0001-9240-291X>

Аннотация. В статье рассматривается актуальность вопроса применения методов науки о данных в геохимии нефти и газа. Для изучения этого вопроса была разработана и реализована на практике методика поиска и сбора научных публикаций за последнее десятилетие из баз данных и их последующего анализа. Выявлен возрастающий интерес к взаимной интеграции этих областей. В работе приведены конкретные примеры найденных публикаций, обозначены ключевые “проблемы” препятствующие широкому внедрению науки о данных в геохимию (необходимость верификации результатов, недостаток квалифицированных специалистов, проблемы с доступом к данным, недоверие к новым подходам) и перспективные идеи для дальнейшего использования методов науки о данных для решения задач органической геохимии (геологические ассистенты, открытые геолого-геохимические базы данных и специализированные цифровые инструменты).

Ключевые слова: органическая геохимия, геохимия нефти и газа, наука о данных, анализ данных, машинное обучение, визуализация данных

Для цитирования: Шевченко Г.А., Большакова М.А. Текущее состояние применения методов науки о данных в геохимии нефти и газа // Вестн. Моск. ун-та. Сер. 4. Геология. 2025. № 1. С. 88–96.

APPLICATIONS OF DATA SCIENCE METHODS IN PETROLEUM GEOCHEMISTRY: CURRENT STATE

Gleb A. Shevchenko^{1✉}, Mariya A. Bolshakova²

¹ Lomonosov Moscow State University, Moscow, Russia; gleb.a.shevchenko@gmail.com ✉

² Lomonosov Moscow State University, Moscow, Russia; m.bolshakova@oilmsu.ru, <https://orcid.org/0000-0001-9240-291X>

Abstract. This paper explores the relevance of applications of Data Science methods in petroleum geochemistry. In order to investigate this topic, a methodology for searching, gathering and analyzing scientific papers published in the last decade was developed and successfully applied. The study reveals a growing interest in integrating Data Science methodology into petroleum geochemistry. The article also presents specific examples of found publication, identifies key “problems” hindering widespread Data Science adoption in geochemistry (including the need for result verification, shortage of qualified specialists, issues regarding access to data and negative sentiment towards new methods), and proposes promising ideas for further utilization of data science methods to tackle challenges presented by organic geochemistry (geological assistants, open access geological and geochemical databases and specialized digital toolkits and software).

Keywords: organic geochemistry, petroleum geochemistry, data science, data analysis, machine learning, data visualization

For citation: Shevchenko G.A., Bolshakova M.A. Applications of Data Science methods in petroleum geochemistry: current state. *Moscow University Geol. Bull.* 2025; 1: 88–96. (In Russ.).

Введение. В последнее десятилетие (с начала 2010-х годов) значительно возросло применение систем искусственного интеллекта во всех областях человеческой жизни. Это коснулось и научной сферы деятельности, и привело к трансформации множества существующих и появлению новых междисциплинарных направлений, которые сочетают в себе естественные и компьютерные науки (биоинформатика, геоинформатика, хемоинформатика и др.), а также к использованию новых, во многом более совершенных методов исследований, позволяющих работать более эффективно.

В данной статье речь пойдет о том, как эта трансформация повлияла на геологию, конкретно на геологию нефти и газа. Актуальность исследования этого вопроса заключается в том, что искусственный интеллект и наука о данных тесно связаны между собой и использование новых методов для анализа нефтегеологической информации и интеграция их в производственные процессы и научную деятельность дают множество преимуществ.

В геологии нефти и газа используются методы многих геологических наук: литология, стратиграфия, геофизика (сейсморазведка и геофизические

исследования скважин), геохимия нефти и газа (органическая геохимия), петрофизика и других. В большинстве опубликованных статей внимание исследователей концентрируется на применении методов науки о данных в нефтегазовой геологии в целом, без разделения на отдельные направления [Шиверский, 2022; Tariq, et al., 2021]. Преимущественно подобные исследования посвящены геофизическим и петрофизическим работам, где широко применяются математические и компьютерные методы для решения практических задач.

Научная новизна работы заключается в том, что в ней будет рассмотрен вопрос применение методов в сфере геохимии нефти и газа.

Целью этой работы является оценка тенденций применения в геохимии (конкретно нефтегазовой/органической) методов науки о данных, посредством обзора научных публикаций.

Для достижения обозначенной цели были поставлены следующие задачи:

- охарактеризовать суть и возможности науки о данных;
- разработать методику для поиска и последующего анализа релевантных публикаций: поиск и выбор источников информации, составление выражения-запроса для их отбора, выборка подходящих под критерии записи;
- собрать и выверить найденную информацию, объединить в единый набор данных и провести его анализ: дедупликация записей, расчет соотношений, визуализация данных для выявления общей картины и трендов;
- описать потенциальные препятствия для более широкого применения в органической геохимии и пути их разрешения;
- обозначить перспективные идеи, которые в дальнейшем могут быть реализованы авторами и другими исследователями.

Наука о данных (англ. Data Science) — это раздел информатики, предметом которого являются данные, все аспекты работы с ними и изучение проблем их сбора, анализа, обработки и представления в цифровой форме. Наука о данных имеет связь не только с искусственным интеллектом, но и со следующими смежными областями [Maslianko, et al., 2021; Sarker, et al., 2021].

Аналитика данных (Data Analytics) включает процессы сбора, обработки, изучения и интерпретации данных для получения практических выводов с использованием традиционных (статистических, эмпирических и логических) методов.

Аналитика больших данных (Big Data) — технологии и инструменты для хранения, обработки и анализа больших по объему и сложных наборов данных.

Сбор и интеллектуальный анализ данных, добыча данных (Data Mining) — извлечение нетривиальной и практически полезной информации из необработанных массивов больших данных.



Рис. 1. Положение науки о данных по отношению к составляющим ее областям знаний в виде диаграммы Венна по [Conway, 2010] с изменениями

Искусственный интеллект (Artificial Intelligence, AI) — алгоритмы, позволяющие выявлять скрытые зависимости и закономерности в данных и в дальнейшем делать предсказания или принимать решения на их основе (можно сказать, что они являются сердцем, центральной составляющей науки о данных); машинное обучение (Machine Learning, ML) — классические, специализированные под конкретные задачи и алгоритмы, и глубокое обучение (Deep Learning, DL) — искусственные нейронные сети.

Инженерия данных (Data Engineering) — построение, обслуживание и управление инфраструктурой для работы с данными.

Визуализация данных (Data Visualization) — создание графиков, диаграмм, карт и других визуальных элементов для понимания, коммуникации результатов анализа и поиска закономерностей.

Важно отметить, что во многих статьях термины «наука о данных» и «машинное обучение» / «искусственный интеллект» (и аналогичные исходно англоязычные термины) зачастую используются в качестве синонимов, хотя, по сути, таковыми не являются (рис. 1) [Sarker, et al., 2021]. Это во многом определяется наибольшим интересом именно к этому направлению в науке о данных, желанием использовать в работах «модные термины» для попадания в поисковые алгоритмы и главенствующим положением этих направлений среди других, которые выполняют подчиненную роль в «хранении, перемещении и обработке данных».

Авторы также будут использовать эти термины в качестве синонимов далее по тексту, поскольку науку о данных можно представить как комбинацию машинного обучения и любой предметной области (в данном случае геохимии нефти и газа) (рис. 1) и термины “Data Science” и “Machine Learning” имеют

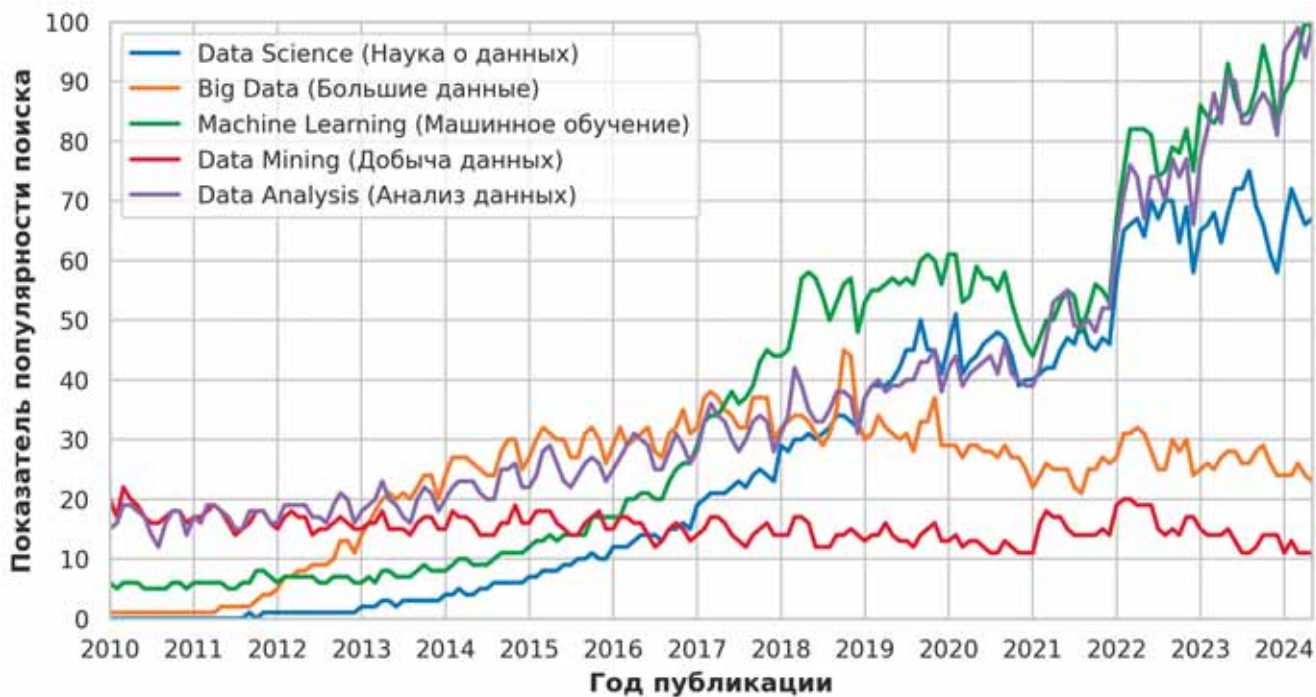


Рис. 2. Мировые тренды популярности поисковых запросов в Google Trends для терминов Data Science и смежных с ней областей за последние 10 лет [Google Trends, 2024]

Таблица 1

Виды представлений информации (по [Maslianko, et al., 2021; Sarker, et al., 2021])

Типы данных		Примеры	Методы обработки и анализа
Структурированные данные: матрица (таблица), массивы		Данные в виде таблиц с определенными строками и столбцами: базы данных (SQL), CSV-файлы, электронные таблицы (Excel, Calc и другие)	Статистический анализ, традиционные методы машинного обучения (регрессия, классификация, кластерный анализ, понижение размерности данных)
Полуструктурированные данные		Данные имеют частичную структуру, содержат теги или маркеры, которые помогают определить их содержание (XML, JSON, HTML, YAML)	Преобразование в структурированный формат или обработка методами для неструктурированных данных
Неструктурированные данные	Изображения	Не имеют определенной структуры и не организованы в виде таблиц	Компьютерное зрение, глубокое обучение — нейронные сети (полносвязанные, сверточные, генеративно-состязательные, автоэнкодеры и другие типы)
	Аудио		
	Тексты	Текстовые документы (отчеты, книги, научные статьи и т. д.)	Обработка естественного языка (Natural Language Processing)
Временные ряды		Последовательность измерений во времени, которое может представляться числом, вектором, а в общем случае — признаковым описанием в данный момент времени: котировки акций на фондовом рынке, показатели погоды, данные датчиков и т. д.	Методы анализа временных рядов (ARIMA, LSTM, скользящее окно и др.)

схожие восходящие тренды популярности (рис. 2) и зачастую употребляются в качестве синонимов.

Типы данных и методы работы с ними. Поскольку данные являются основой любого исследования и предметом изучения Data Science, необходимо обозначить их основные типы, которые в том или ином виде встречаются в любой исследовательской работе, и методы, применяемые для их обработки и анализа (табл. 1).

Методика исследования. Для оценки заинтересованности научного сообщества в вопросе симбиоза науки о данных и органической геохимии были проанализированы крупнейшие базы научных публикаций с открытыми метаданными (для того, чтобы обеспечить прозрачность и повторяемость получаемых результатов) и отсутствием ограничений на максимальное количество выдаваемых в результатах записей [Gusenbauer, 2023].

Таблица 2

Поисковые теги, разделенные по принадлежности «компонентам» (тематикам)

Компоненты Язык	Геохимические	Цифровые
Английский	“organic matter”, “petroleum geochemistry”, “organic geochemistry”, “pyrolysis”, “source rocks”, “total organic carbon”, «reservoir geochemistry», «biomarker», «kinetics», «biodegradation», «geochemical indices», «surface geochemistry»	“artificial intelligence”, “machine learning”, “deep learning”, “data science”, “data analysis”, “database”, “data visualization”, “big data”, “neural networks”
Русский	«органическое вещество», «геохимия нефти и газа», «органическая геохимия», «пиролиз», «нефтегазоматеринские толщи», «органический углерод», «резервуарная геохимия», «биомаркер», «биодegradация», «геохимические коэффициенты»	«искусственный интеллект», «наука о данных», «машинное обучение», «глубокое обучение», «нейронные сети», «анализ данных», «база данных», «визуализация данных», «большие данные»

Международные: OpenAlex (включает Directory of Open Access Journals, CrossRef, Microsoft Academic Graph, Research Organization Registry, Unpaywall) [Priem, et al, 2022]. Web of Science и Scopus не использовались по причине отсутствия к ним открытого доступа, их коммерческой природы и существенного пересечения метаданных с OpenAlex, что позволяет использовать только ее для наукометрической оценки [Лутай, 2022; Culbert, et al., 2024];

Российские: eLIBRARY.ru (РИНЦ).

В процессе поиска релевантных статей использовались следующие **критерии**.

- 1) Использовались теги, которые можно разделить на составляющие «компоненты» (табл. 2).
- 2) Даты публикации: последние 10 лет (2014–2024).
- 3) Поиск осуществлялся в следующих поисковых полях: название публикации и аннотация. Поиск по ключевым словам не использовался по причине сравнительной ненадежности, поскольку максимальное количество слов на одну статью ограничено и эти слова зачастую могут охватывать не все тематики статьи, в отличие от названий и аннотаций, где всегда заключен основной смысл.
- 4) Производился отбор по тематикам и областям науки (field) с последующей дополнительной фильтрацией выборок по тематикам и подобластям (subfield):

eLIBRARY.ru: Геология, Информатика, Кибернетика (Тематика); Геология (Тематика);

OpenAlex: Engineering, Computer Science, Earth and Planetary Sciences, Chemistry (Field); Artificial Intelligence, Mechanics of Materials, Analytical Chemistry, Geochemistry and Petrology, Ocean Engineering, Geology, Information Systems (Subfield).

В целом, **последовательность шагов** по созданию необходимого для анализа набора данных следующая:

- 1) Извлечение данных: экспорт записей из обозначенных выше баз, используя встроенные в них средства для скачивания и скрейпинг локально сохраненных веб-страниц с использованием скрипта на языке Python.

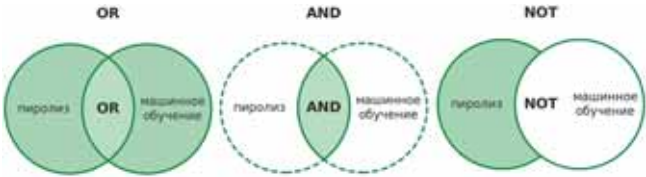


Рис. 3. Логические операторы, которые используются в поисковых системах при построении выражений-запросов: OR — записи, включающие хотя бы один из тегов; AND — записи, включающие оба тега; NOT — записи, включающие один тег и не включающие другой

Записи фильтровались с использованием комбинации различных тегов (при этом также учитывалась морфология слова), складывающихся из двух компонент по следующему принципу: (“biomarker” OR «органическая геохимия» OR ...) AND (“data science” OR «машинное обучение» OR ...) (рис. 3).

- 2) Очистка и объединение данных: приведение их к табличному формату (актуально для веб-страниц), унификация названий колонок метаданных и удаление лишних, дедупликация записей в пределах каждого из наборов и после объединения наборов из разных источников на основе названий.

- 3) Получение дополнительной информации о данных: подсчет встречаемости тегов в поисковых полях и определение количества общих записей между базами данных.

- 4) Визуализация результатов: представление результатов с использованием диаграмм Венна, линейных диаграмм трендов, столбчатых диаграмм распределения по годам и частот встречаемости тегов.

Данный литературный обзор и сам является примером использования методики одного из направлений науки о данных на практике, поскольку в ходе его выполнения осуществлена работа по сбору и анализу данных с визуализацией результатов и дальнейшей формулировкой выводов.

Результаты и обсуждение. Характеристики собранных данных по состоянию на конец мая 2024 г. представлены ниже в табл. 3.

Пересечения баз данных (общих записей, встречающихся в обеих базах) имеют вид, представлен-

Таблица 3

Характеристика собранных наборов данных

База данных \ Включаемые термины	«Геохимические» AND «Цифровые», шт	«Геохимические», шт	Доля «цифровой» части в результатах
OpenAlex	2661	47461	0,056 (5,6 %)
eLIBRARY.ru	556	12636	0,044 (4,4 %)

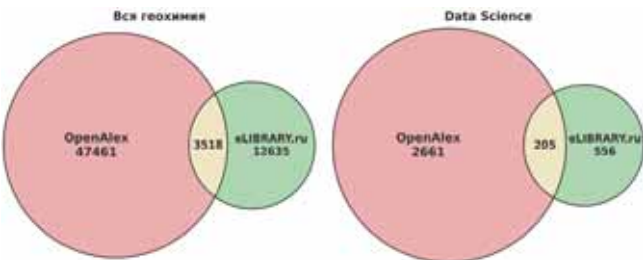


Рис. 4. Диаграммы Венна, показывающие количество записей, присутствующих в каждом из наборов данных и общих для обеих баз записей (табл. 3)

ный на рис. 4. Для расчета величины перекрытия использовались только названия статей, так как не удалось осуществить экспорт из eLIBRARY.ru поля DOI (цифрового идентификатора объекта), который бы позволил однозначно определить и вычислить количество дубликатов.

Диаграммы количества публикаций как по геохимической тематике в целом, так и по тематике науки о данных, построенные по данным из обоих источников, представлены на рис. 5. По ним хорошо видно, что количество публикаций возрастает от года к году, так же, как и доля статей, в которых используется методика науки о данных.

На рис. 6 представлена встречаемость используемых терминов. Видно, что теги “machine learning”,

“database”, “neural networks”, “geochemical”, “biomarker” характеризуется наибольшей встречаемостью.

Описание конкретных примеров. Из собранного набора мы выбрали некоторые примеры статей, которые иллюстрируют возможности и использование применительно к геохимическому направлению методов науки о данных. Публикации из выборки можно разделить на две группы по целям и решаемым задачам.

Анализ с помощью методов науки о данных — целью исследования является решение геохимических задач, использующих эти методы (табл. 4).

Цифровые инструменты на языке Python — целью такого исследования является создание инструмента для упрощения процесса анализа данных методами науки о данных, описание этого процесса и схемы работы решения:

1) Geochemistry π [ZhangZhou, et al., 2024] — Python-фреймворк для геохимических исследований с использованием автоматического машинного обучения (AutoML) и табличных геохимических данных;

2) GeoPyTool [Yu, et al., 2018] — программное обеспечение для визуализации и анализа геохимических и структурных данных;

3) pyrolite [Williams, et al., 2020] — набор инструментов для визуализации, обработки и преобразования геохимических данных.

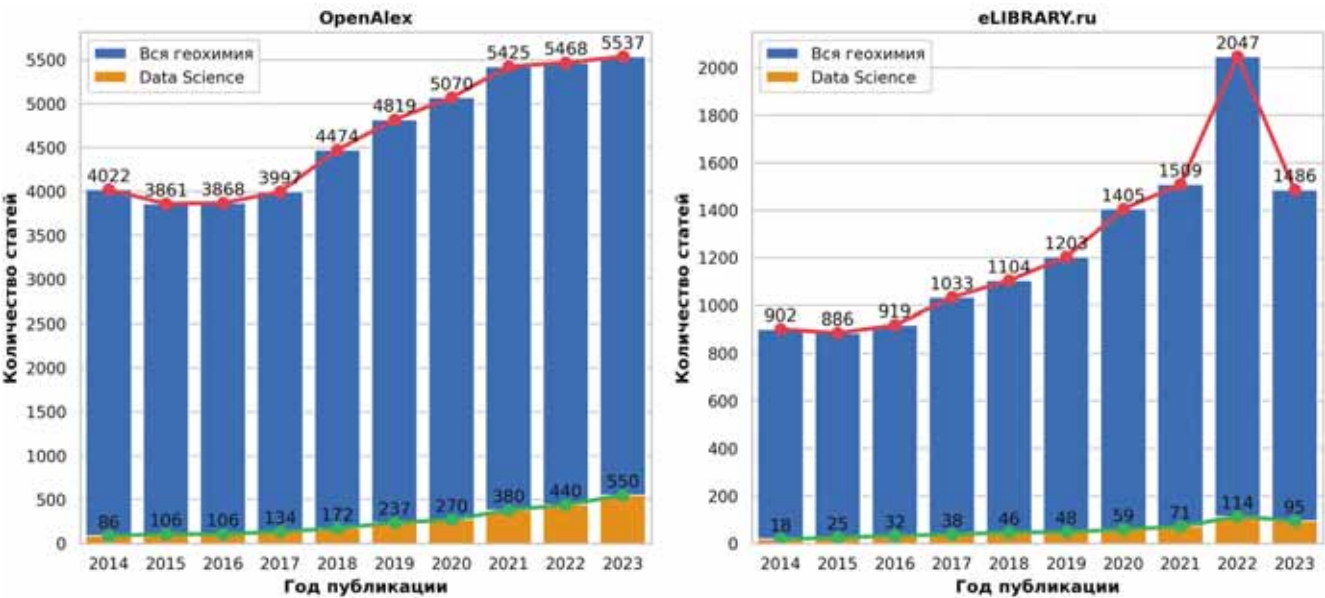


Рис. 5. Столбчатая диаграмма с накоплением по годам количества публикуемых статей, полученных по результатам анализа публикаций из двух баз данных

Таблица 4

Анализ материалов статей с помощью методов науки о данных

Статья	Краткое описание	Используемые методы	Результаты
Prediction of TOC content ... (Sun, et al., 2023)	Сравнительный анализ машинного обучения с традиционными эмпирическими методами для предсказания значений ТОС в породах с высоким содержанием органического вещества, оценка важности различных каротажных кривых для отбора наиболее важных признаков при прогнозировании ТОС.	Машинное обучение (случайный лес, дерево решений, метод опорных векторов и градиентный бустинг XGBoost).	Модель случайного леса продемонстрировала наивысшую точность прогнозирования ТОС ($R^2 = 0,915$), превосходя XGBoost ($R^2 = 0,847$), метод опорных векторов ($R^2 = 0,762$) и традиционные методы (R^2 от 0,003 до 0,137).
Intelligent geochemical interpretation ... (Su, et al., 2024)	Автоматическая интерпретация изображений масс-хроматограмм с использованием нейронных сетей для определения характеристик биомаркеров нефти и проведения корреляции нефть-НГМТ.	Сверточные нейронные сети (CNN), факторный анализ для понижения размерности данных.	CNN успешно интерпретирует масс-хроматограммы, определяя зрелость органического вещества (ОВ) и тип материнской породы. Точность сопоставима с ручной интерпретацией, но автоматизация значительно ускоряет процесс и позволяет извлекать информацию из масс-хроматограмм, которая недоступна при традиционной интерпретации.
Geochemical biodegraded oil classification ... (Bispo-Silva, et al., 2023)	Определение степени биodeградации нефтей по изображениям хроматограмм с использованием сверточных нейронных сетей, сравнительный анализ нескольких алгоритмов машинного обучения.	Сверточные нейронные сети (CNN), машинное обучение (наивный байесовский классификатор, случайный лес, дерево решений и логистическая регрессия).	CNN успешно классифицирует хроматограммы на биodeградированные и небиodeградированные с accuracy = 97,6 %. Нейронные сети показали наилучшую accuracy = 96,7 % среди всех тестируемых алгоритмов машинного обучения.
Выявление особенностей ... (Осипов и др., 2022)	Определение различий органического вещества двух нефтегазоматеринских толщ (НГМТ), осуществление корреляции нефть-НГМТ, сравнение методов геохимического (традиционного) и статистического анализов.	Машинное обучение (случайный лес, k-ближайших соседей, анализ главных компонент для понижения размерности).	Результаты статистического анализа хорошо согласуются с результатами традиционного геохимического анализа. Поэтому статистический анализ (включает методы машинного обучения) является эффективным инструментом для выявления различий между НГМТ и корреляции нефть-НГМТ.
Denoising of Geochemical Data ... (Zhang, et al., 2024)	Исследование возможности очистки данных региональных геохимических исследований от шумов.	Нейронные сети (автоэнкодеры).	Автоэнкодеры эффективно уменьшают шум, что проявляется в уменьшении эффекта саморodka на вариограммах. Определена линейная зависимость между степенью снижения шума и производительностью автоэнкодера. Производительность также связана с показателями качества данных, такими как точность, прецизионность и F-статистика (ANOVA). Определен оптимальный размер кодирующего слоя (11 узлов), который обеспечивает максимальное снижение шума без удаления значимых геохимических аномалий.
DATA MINING IN ORGANIC ... (Torres, et al., 2022)	Разработка методологии анализа геохимических данных нефти с использованием интеллектуального анализа данных и машинного обучения для классификации происхождения нефтей.	Машинное обучения (случайный лес, дерево решений и логистическая регрессия, искусственные нейронные сети), иерархический кластеризация, кластеризация методом k-средних, многомерное шкалирование (MDS) для снижения размерности данных.	Из 60 исходных атрибутов с помощью MDS был выделен оптимальный набор из 26 переменных. Классификация 200 образцов нефти с помощью ML-алгоритмов показала следующие accuracy: 92.50 % для дерева решений, 95.00 % для случайного леса и 87.50 % для искусственной нейронной сети. Предложенный метод сопоставим по точности с результатами, полученными традиционными методами.

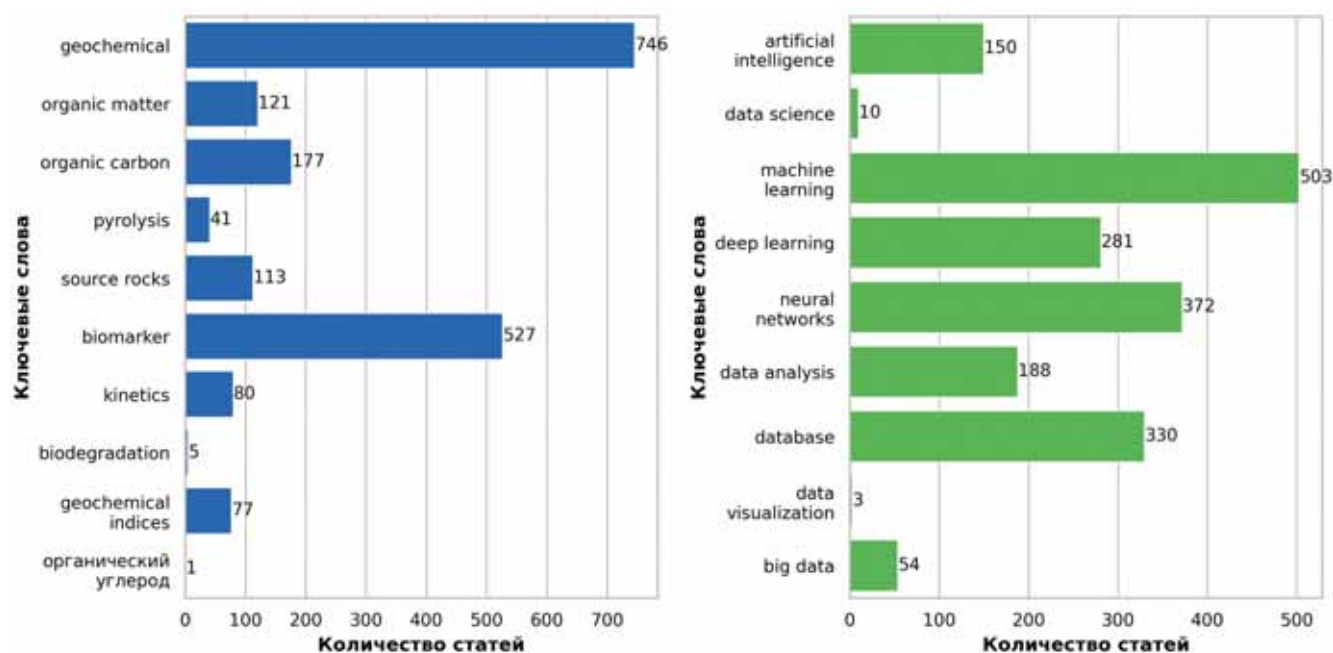


Рис. 6. Диаграмма суммарной встречаемости тегов в названиях и аннотациях

Факторы, мешающие расширению использования методов науки о данных применительно к геологии. Исходя из изложенного выше напрашивается вывод, что включение методов науки о данных должно принести качественные улучшения в исследовательские процессы. Однако полученные в результате исследования цифры (табл. 3) свидетельствуют о том, что существуют некие препятствия для широкого применения этих методов.

1. Проблема интерпретируемости моделей и верификации результатов. Модели машинного обучения могут давать точные прогнозы, но быть при этом «черными ящиками», т.е. не объяснять, почему они пришли к такому выводу, так как причинно-следственные связи в сложных моделях не имеют выражения в явном виде. Это затрудняет интерпретацию и верификацию полученных результатов, что, зачастую, и является самой сложной задачей в процессе исследования, решить которую может только опытный специалист. В связи с этим возникает вторая проблема.

2. Скептическое отношение к новым подходам. Исследования в области геохимии нефти и газа пока могут быть больше ориентированы на применение традиционных, чем переход на использование новых цифровых методов, что можно кратко охарактеризовать выражением «не надо ломать то, что уже работает правильно». Эксперты могут быть не осведомлены о возможностях этих методов, им может не хватать должных навыков для их корректного применения, или это может быть связано с проблемой интерпретируемости результатов.

3. Ограниченный доступ к данным. Большое количество геолого-геохимических данных является собственностью компаний или государства и может быть недоступно для проведения исследований. Так

в Российской Федерации такая информация может предоставляться только по запросу или представлять собой коммерческую и государственную тайну. Это приводит к ограничению размера выборок и следующей проблеме.

4. Объем данных для обучения и особенности «геологических объектов». Модели машинного обучения во многом основаны на принципах математической статистики. В связи с этим, необходимо достаточное количество данных для обучения и получения точных результатов. Это не всегда достижимо, поскольку когда идет речь о геологических данных, то количества «реальных» замеров для получения репрезентативной выборки может просто не хватать. Поэтому существуют специальные техники, которые, например, генерируют синтетические «промежуточные» данные между известными «реальными», что, однако, означает дополнительные вычисления и временные затраты на их реализацию и выполнение. В то же время каждый геологический объект или процесс является по-своему уникальным и индивидуальным, что зачастую плохо поддается однозначной классификации и интерпретации даже при его изучении учеными-специалистами с использованием классических методов. Создание единой модели для нескольких таких объектов не всегда возможно по причине отличающихся параметров распределений, описывающих интересующие модельный свойства, например, две разновозрастные нефтегазоматеринские толщи с принципиально разными параметрами генерационным потенциалом.

5. Высокая стоимость оборудования и недостаток вычислительных мощностей для обучения и эксплуатации моделей. Большая часть методов машинного обучения и даже простых нейронных сетей на сегод-

нящий момент не требует много вычислительных ресурсов, и они способны быстро исполняться практически на любом вычислительном устройстве, но не для сложных мультимодальных или полнотекстовых моделей, которые требуют достаточно дорогого оборудования (видеокарта или ИИ-ускоритель) для их локального использования или привязки к облачным сервисам для их удаленного исполнения. В случаях когда требуется работать с большим объемом данных, необходимых для обучения самих моделей, то задача только усложняется, поскольку возрастают требования к размерам хранилищ информации, что требует значительных инвестиций в оборудование.

6. Сложность решаемой задачи. При решении простой задачи, не требующей большого количества данных, автоматически отпадает необходимость в реализации специализированной системы хранения и обработки данных вместо использования, например, обычных электронных книг Excel в виду неоправданных временных затрат. То же самое можно сказать и про применение алгоритмов машинного обучения вместо простых, сугубо алгоритмических методов или классических методов анализа данных.

7. Новизна области исследования и нехватка квалифицированных специалистов. Применение методов науки о данных при работе с геолого-геохимическими данными — активно развивающаяся область (что видно исходя из трендов роста публикаций), поэтому многие исследования могут быть еще не опубликованы или находиться в стадии разработки. Кроме того, проведение подобных исследований и реализация новых цифровых решений требует наличие глубокими знаниями как в предметной области — геологии, так и в других составляющих науку о данных областях — информатике и математике (рис. 1). В настоящий момент эта проблема активно решается созданием междисциплинарных программ для подготовки специалистов, имеющих необходимую экспертизу в обеих областях.

Перспективы использования методов науки о данных в органической геохимии.

1. Геологические ассистенты. Адаптация (тонкая настройка) большой языковой или мультимодальной (GPT, Mistral, Llama и прочие) модели для решения геологических (и геохимических) задач, способных отвечать на вопросы, помогать пользователю в обработке данных, генерировать графические материалы [Cheng, et al., 2023; Lin, et al., 2023].

2. Геолого-геохимические базы данных. Создание единых, обезличенных баз данных с результатами геохимических исследований и свободным доступом: проектирование структуры баз, организация сбора и систематизации данных из разных источников, их администрирование, интеграция с другими источниками геологической информации, предоставления внешнего доступа для пользователей и разработчиков через API или веб-интерфейс [Осипов и др., 2022; Farrell, et al., 2021; Wyborn, et al., 2020].

3. Цифровые инструменты. Инструменты для решения рутинных и узкоспециализированных задач органической геохимии, такие как обработка определение степени биodeградации нефтей по хроматограммам, полуавтоматическая интерпретация пиков; определение условий осадконакопления, источников и типов ОВ, степени катагенетической преобразованности, литологического состава вмещающей НГМТ, на основе геохимических коэффициентов и индексов; визуализация карт и специализированных геохимических диаграмм; автоматическая интерпретация результатов пиролиза с использованием геохимических диаграмм.

Заключение. Проведенное исследование показало, что наблюдается рост числа статей, совместно упоминающих науку о данных и геохимию нефти и газа, что свидетельствует о том, что методы науки о данных на данный момент находят свое применение в этой области. Однако, в целом, востребованность подобного рода исследований пока не так высока, что хорошо отражают вычисленные доли «цифровой» статей среди всех результатов: 5.6 (OpenAlex) и 4.4 (eLIBRARY.ru) %.

Обозначены 7 проблем, препятствующих использованию методики науки о данных абсолютно во всех исследовательских проектах, связанных не только с геохимией, но и геологией нефти и газа в целом.

В работе предложены 3 варианта использования методов науки о данных.

Включение методов науки о данных в практику геохимии нефти и газа является, по мнению авторов, неизбежной и необходимой тенденцией для поддержания актуальности и конкурентоспособности геологии нефти и газа в целом среди других естественных наук. Эта тенденция, пусть медленно, но реализуется на практике множеством исследователей. Сейчас создание подобных инструментов является задачей значительно более простой и актуальной, чем когда бы то ни было до этого в силу наличия таких технологий как язык программирования Python и его огромной экосистемы библиотек для работы с любыми типами данных в во всех возможных форматах, любыми возможными методами и наличия открытого доступа к огромному количеству обучающих материалов в интернете.

Авторы также хотят отметить, что предложенная методика поиска и сбора материалов не является совершенной — ими могла быть найдена лишь часть отвечающих критериям поиска работ, поэтому собранный набор статей и полученные результаты не является исчерпывающим. Так, например, могут быть улучшены следующие аспекты поиска и обработки данных:

1) использование при поиске публикаций других открытых, но ограничивающих массовый экспорт записей баз данных [Gusenbauer, 2023]: Bielefeld Academic Search Engine (BASE), GeoRef, CORE

и OpenAIRE, — это должно позитивно сказаться на полноте выборки;

2) добавление дополнительных слов в поисковой набор тегов, создание набора тегов для других языков кроме русского и английского;

3) осуществление дедупликации статей между базами данных, используя другие поля метаданных, такие как имена авторов и DOI.

СПИСОК ЛИТЕРАТУРЫ

Лутай А.В., Любушко Е.Э. Сравнение качества метаданных в БД CrossRef, Lens, OpenAlex, Scopus, Semantic Scholar, Web of Science Core Collection. Российский фонд фундаментальных исследований (РФФИ). 2022. URL: https://podpiska.rfbr.ru/storage/reports2021/2022_meta_quality.html (дата обращения: 28.05.2024).

Осипов К.О., Абля Э.А., Сауткин Р.С. и др. Выявление особенностей органического вещества нефтей и нефтегазоматеринских толщ путем сопоставления результатов геохимического анализа со статистическим анализом, основанным на методах машинного обучения (на примере одного из месторождений Западно-Сибирского нефтегазового бассейна) // Георесурсы. 2022. Т. 24. № 2. С. 217–229.

Шиверский Г.В., Кривошеков С.Н. Перспективы применения методов искусственного интеллекта в нефтегазовой геологии // Журнал магистров. 2022. № 2. С. 57–67.

Bispo-Silva S., Oliveira C.J., De Alemar Barberes G. Geochemical biodegraded oil classification using a machine learning approach // Geosciences. 2023. Vol. 13. N 11. P. 321.

Cheng D., Zhang T., He Z., et al. K2: A Foundation Language Model for Geoscience Knowledge Understanding and Utilization // arXiv e-prints. 2023. URL: <https://doi.org/10.48550/arxiv.2306.05064>

Conway D. The Data Science Venn Diagram. Drew Conway Data Consulting. 2010. URL: <http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram> (дата обращения: 31.05.2024).

Culbert J., Hobert A., Jahn N., et al. Reference Coverage Analysis of OpenAlex compared to Web of Science and Scopus // arXiv e-prints. 2024. URL: <https://doi.org/10.48550/arXiv.2401.16359>

Farrell Ú.C., Samawi R., Anjanappa S., et al. The Sedimentary Geochemistry and Paleoenvironments Project // Geobiology. 2021. Vol. 19. N 6. P. 545–556.

Google Colab. URL: <https://colab.research.google.com/drive/1oULyOqrP90-SVIEy1RRM2SNHgAKpHb?usp=sharing> (дата обращения: 31.05.2024).

Gusenbauer M. A free online guide to researchers' best search options // Nature. 2023. Vol. 615. P. 586.

Lin Z., Deng C., Zhou L., et al. GeoGalactica: A Scientific Large Language Model in Geoscience // arXiv e-prints. 2023. URL: <https://doi.org/10.48550/arXiv.2401.00434>

Данные и код. Для создания иллюстраций и обработки данных использовалась интерактивная облачная платформа Google Colab для работы с кодом на Python. Блокнот с кодом и пояснениями находится в открытом доступе [Google Colab, 2024].

Благодарности. Авторы статьи признательны рецензентам за потраченное время и ценные замечания.

Maslianko P., Sielskiy Y. Data Science — definition and structural representation // System Research & Information Technologies. 2021. N 1. P. 61–78.

Priem J., Piwowar H., Orr R. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts // arXiv e-prints. 2022. URL: <https://doi.org/10.48550/arXiv.2205.01833>

Sarker I.H. Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective // SN Computer Science. 2021. Vol. 2. N 5. URL: <https://doi.org/10.1007/s42979-021-00765-8>

Su K., Lu J., Yu J., et al. Intelligent geochemical interpretation of mass chromatograms: Based on convolution neural network // Petroleum Science. 2024. Vol. 21, N 2. P. 752–764.

Sun J., Dang W., Wang F., et al. Prediction of TOC content in Organic-Rich shale using machine learning algorithms: comparative study of random forest, Support Vector Machine, and XGBOOST // Energies (Basel). 2023. Vol. 16, N 10. P. 4159.

Tariq Z., Aljawad M.S., Hasan A., et al. A systematic review of data science and machine learning applications to the oil and gas industry // Journal of Petroleum Exploration and Production Technology. 2021. Vol. 11, N 12. P. 4339–4374.

Torres S.B., De Oliveira Matias Í., De Araújo Ponte F.F., et al. Data mining in organic geochemistry: case study in Potiguar basin // Geociências. 2022. Vol. 41, N 1. P. 105–114.

Williams M.J., Schoneveld L., Mao Y., et al. pyrolite: Python for geochemistry // Journal of Open Source Software. 2020. Vol. 5. N 50. P. 2314.

Wyborn L., Lehnert K.A. OneGeochemistry: Creating a global FAIR-Way to access and share geochemical data // Goldschmidt Abstracts. 2020. URL: <https://doi.org/10.46427/gold2020.2910>

Yu Q.-Y., Bagas L., Yang P.-H., et al. GeoPyTool: a cross-platform software solution for common geological calculations and plots // Geoscience Frontiers. 2019. Vol. 10. N 4. P. 1437–1447.

Zhang S.E., Bourdeau J.E., Nwaila G.T., et al. Denoising of Geochemical Data using Deep Learning—Implications for Regional Surveys // Natural Resources Research. 2024. Vol. 33. P. 495–520.

Zhangzhou J., He C., Sun J., et al. Geochemistry π: Automated machine learning Python framework for tabular data // Geochemistry, Geophysics, Geosystems. 2024. Vol. 25. N 1. P. e2023GC011324.

Статья поступила в редакцию 02.06.2024, одобрена после рецензирования 30.10.2024, принята к публикации 28.02.2025