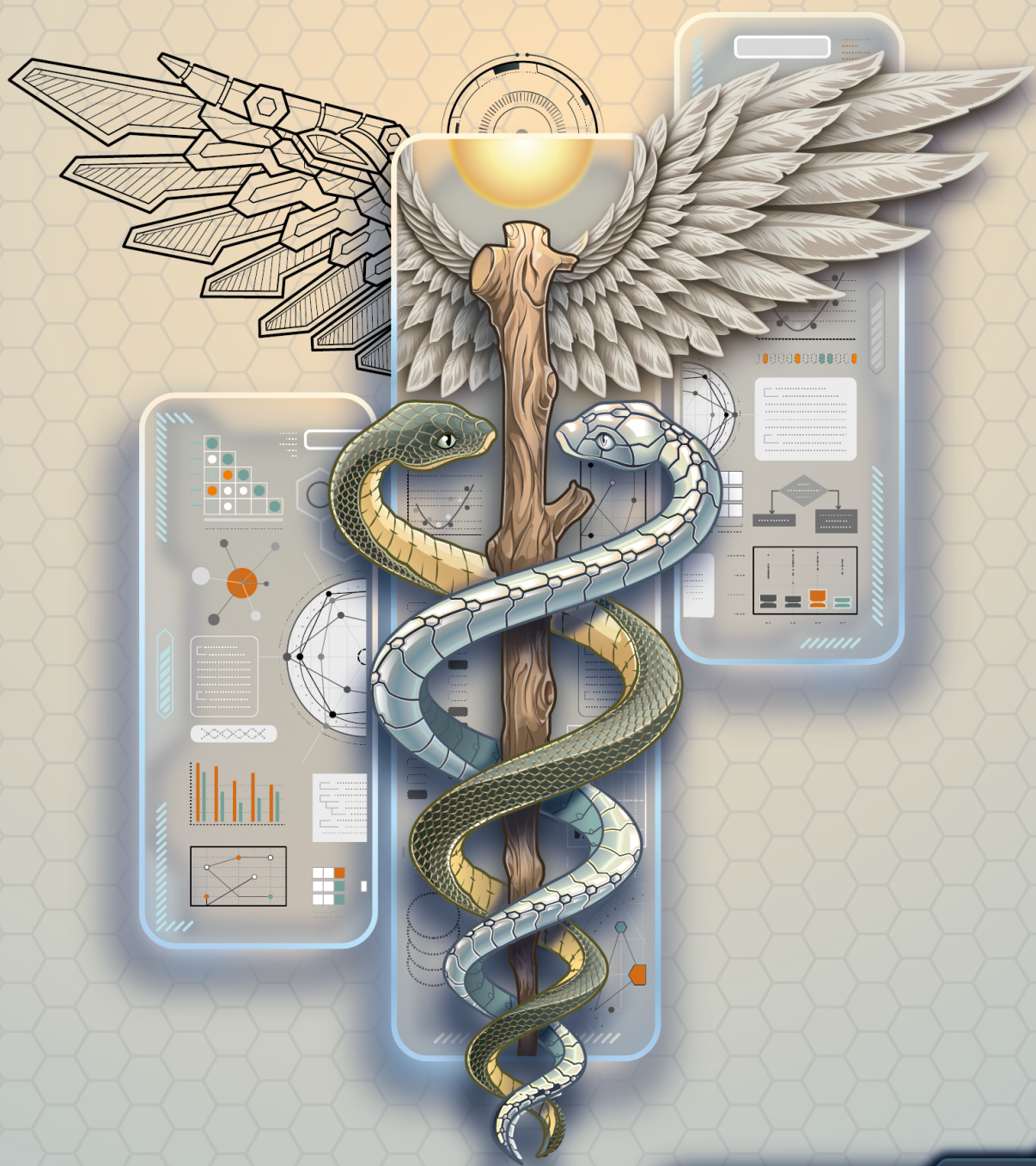


НАУКА О ДАННЫХ И ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ В МЕДИЦИНЕ



А.Ю. КУЗЬМЕНКОВ
А.Г. ВИНОГРАДОВА
И.В. ТРУШИН
А.А. АВРАМЕНКО
М.Ю. КУЗЬМЕНКОВ

**А.Ю. КУЗЬМЕНКОВ
А.Г. ВИНОГРАДОВА
И.В. ТРУШИН
А.А. АВРАМЕНКО
М.Ю. КУЗЬМЕНКОВ**

НАУКА О ДАННЫХ И ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ В МЕДИЦИНЕ

УДК 004.8:61
ББК 5с51
Н 34

Н 34 **Наука о данных и искусственный интеллект в медицине.** / А.Ю. Кузьменков, А.Г. Виноградова, И.В. Трушин, А.А. Авраменко, М.Ю. Кузьменков — М.: Е-нот, 2026. — 744 с.

ISBN 978-5-906023-48-3

«Наука о данных и искусственный интеллект в медицине» — уникальное практическое руководство, объединяющее современные подходы анализа больших данных и искусственного интеллекта с прикладными аспектами здравоохранения. Эта книга охватывает весь цикл работы с биомедицинскими данными: от планирования клинических исследований и сбора информации до углубленного статистического анализа и публикации результатов. Читатель познакомится с основами программирования на языке R, различными системами управления базами данных и алгоритмами машинного обучения, а также получит возможность отработать полученные навыки на реальных наборах биомедицинских данных. Издание станет незаменимым помощником как для опытных лечащих врачей и ученых-исследователей, так и для студентов вузов и специалистов из разных областей, желающих овладеть новыми компетенциями.

Рецензенты:

Колбин Алексей Сергеевич — доктор медицинских наук, профессор, заведующий кафедрой клинической фармакологии и доказательной медицины ФГБОУ ВО «Первый Санкт-Петербургский государственный медицинский университет имени академика И.П. Павлова» Министерства здравоохранения Российской Федерации

Середа Андрей Петрович — доктор медицинских наук, профессор кафедры восстановительной медицины, лечебной физкультуры и спортивной медицины, курортологии и физиотерапии Академии постдипломного образования ФГБУ «Федеральный научно-клинический центр специализированных видов медицинской помощи и медицинских технологий Федерального медико-биологического агентства» Российской Федерации

Балыкина Юлия Ефимовна — кандидат физико-математических наук, доцент кафедры математического моделирования энергетических систем ФГБОУ ВО «Санкт-Петербургский государственный университет»

УДК 004.8:61
ББК 5с51
Н 34

ISBN 978-5-906023-48-3

© ФГБОУ ВО СГМУ Минздрава России, 2025

***«Большое дерево вырастает из маленького,
девятиэтажная башня начинает строиться из горстки земли,
путешествие в тысячу ли начинается с одного шага...»***

Лао-цзы

«...изучение науки о данных начинается со следующей страницы.»

Коллектив авторов

Оглавление

Оглавление	5
Об авторах	11
Предисловие	12
Список сокращений	14
Глава 1. Особенности медицинских исследований	16
1.1. Основные понятия доказательной медицины	16
1.1.1. Постановка клинически значимого вопроса	17
1.1.2. Поиск наилучших доказательств	18
1.1.3. Критическая оценка доказательств	18
1.1.4. Применение доказательств	21
1.1.5. Оценка эффективности предложенных подходов на основе принципов ДМ	22
1.2. Планирование биомедицинских исследований	22
1.2.1. Дизайн исследования	23
1.2.2. Описательные исследования	24
1.2.3. Аналитические исследования	25
1.2.4. Контроль	28
1.2.5. Рандомизация	28
1.2.6. Ослепление	30
1.2.7. Типичные схемы дизайна биомедицинских исследований	30
1.3. Клинические исследования	32
1.3.1. Фазы КИ	32
1.3.2. Клинический вопрос и уровни доказательности	34
1.3.3. Регламентирующие документы КИ	36
1.3.4. Роли участников КИ	36
1.3.5. Конечные точки КИ	36
1.3.6. Критерии включения и исключения	38
1.3.7. Систематические ошибки в КИ	38
1.3.8. Принципы проведения КИ	39
1.4. Систематический обзор и метаанализ	41
1.4.1. Систематический обзор	42
1.4.2. Метаанализ	45
1.4.3. Публикационное смещение	48
Глава 2. Основы программирования	50
2.1. Язык программирования R	50
2.1.1. RStudio	51

2.1.2. Типы данных.....	53
2.1.3. Переменные	55
2.1.4. Математические операторы.....	56
2.1.5. Операторы сравнения и логические операторы	57
2.1.6. Операторы проверки на значения	58
2.2. Структуры данных, функции, циклы.....	58
2.2.1. Функции	58
2.2.2. Пакеты.....	61
2.2.3. Структуры данных.....	62
2.2.4. Встроенные наборы данных.....	79
2.2.5. Ветвления и условные переходы	81
2.2.6. Циклы	83
2.3. Особенности подготовки табличных данных.....	85
2.3.1. Рекомендации по подготовке таблиц для чтения в R.....	85
2.3.2. Манипуляции с данными	90
2.3.3. Работа с датами	105
2.3.4. Работа с числовыми данными	111
2.3.5. Работа со строками	115
2.3.6. Грамматика манипуляций с данными с использованием dplyr	123
2.3.7. Конвейер обработки (пайплайн)	131
2.3.8. Работа с данными с использованием data.table	133
2.3.9. Особенности работы с data.table, data.frame и tibble.....	140
2.3.10. Сравнение dplyr и data.table.....	142
2.3.11. Объединение таблиц.....	143
2.3.12. Соединение таблиц	144
2.3.13. Разворот таблиц	148
2.3.14. Сохранение данных.....	150
2.4. Валидация и очистка данных	151
2.4.1. Концепция «опрятных» данных	151
2.4.2. Валидация данных.....	171
2.5. Получение данных из внешних источников.....	178
2.5.1. Загрузка таблиц из баз данных	178
2.5.2. Загрузка данных из Интернета.....	182
2.5.3. Отправка HTTP-запросов и работа с API.....	183
2.5.4. Работа с HTTP-запросами в R с использованием httr2	189
2.5.5. Работа с JSON.....	199
2.5.6. Формат XML и HTML.....	211
2.5.7. Парсинг страниц в Интернете	221
2.5.8. Веб-скрэпинг с R Selenium	229
2.6. Сохранение результатов работы.....	235
2.6.1. Подготовка данных	235

2.6.2. Сохранение таблиц	238
2.6.3. Сохранение изображений	242
Глава 3. Системы управления базами данных	245
3.1. Понятие о СУБД.....	245
3.1.1. Типы СУБД.....	245
3.1.2. Реляционные СУБД. Основные понятия	254
3.1.3. Нормализация и денормализация данных	262
3.2. PostgreSQL.....	265
3.2.1. Основные объекты и типы данных в PostgreSQL.....	266
3.2.2. Операции с различными типами данных.....	270
3.3. Команды SQL DDL.....	272
3.3.1. CREATE	272
3.3.2. ALTER и RENAME.....	275
3.3.3. DROP и TRUNCATE	276
3.4. Команды SQL DML	277
3.4.1. INSERT, UPDATE и DELETE.....	277
3.4.2. SELECT	278
3.4.3. Вспомогательные операторы команд DML	279
3.4.4. Вспомогательные функции команд DML	286
3.4.5. Подзапросы.....	290
3.5. Продвинутое выборки	293
3.5.1. Материализованные представления.....	293
3.5.2. Наборы группирования	294
3.5.3. Запросы WITH	298
3.5.4. Оконные функции.....	300
3.5.5. Приемы для оптимизации запросов	305
Глава 4. Визуализация биомедицинских данных	315
4.1. Введение в ggplot2.....	317
4.1.1. Грамматика графических элементов.....	317
4.1.2. Основные компоненты ggplot2.....	318
4.1.3. Сохранение графика.....	331
4.2. Основные подходы к визуализации данных.....	332
4.2.1. Одномерные диаграммы для количественных переменных.....	332
4.2.2. Одномерные диаграммы для качественных переменных	335
4.2.3. Диаграммы для сравнения двух количественных переменных	339
4.2.4. Диаграммы для сравнения двух качественных переменных.....	342
4.2.5. Диаграммы для сравнения качественных и количественных переменных	344

4.3. Продвинутое способы визуализации	349
4.3.1. Лепестковые диаграммы	350
4.3.2. Диаграмма Санкей.....	351
4.3.3. Тепловая карта.....	352
4.3.3. Матрица корреляции	354
4.4. Пакеты серии grammar of graphics	356
Глава 5. Общие вопросы статистического анализа	358
5.1. Выборка, переменные и их описание	358
5.1.1. Выборка и генеральная совокупность	358
5.1.2. Переменные и шкалы	359
5.1.3. Описание переменных.....	362
5.1.4. Понятие о доверительном интервале.....	386
5.2. Разница эффектов и статистические гипотезы	394
5.2.1. Понятие о разнице эффектов.....	394
5.2.2. Понятие о статистических гипотезах.....	398
5.2.3. Понятие о значении p , критическом значении и статистической значимости.....	401
5.2.4. «Параметрические», «непараметрические» и перестановочные тесты.....	408
Глава 6. Частные вопросы статистического анализа.....	410
6.1. Статистические тесты для качественных данных.....	410
6.1.1. Тесты семейства χ^2 и критерий согласия Пирсона	411
6.1.2. Точный критерий Фишера.....	415
6.1.3. Критерий МакНемара для зависимых групп	417
6.1.4. Алгоритм выбора теста для качественных данных.....	420
6.2. Статистические тесты для количественных данных	421
6.2.1. Нормальное распределение	421
6.2.2. Распределение, отличное от нормального.....	432
6.2.3. Количественные сравнения нескольких групп	435
6.2.4. Алгоритм выбора теста для количественных данных.....	445
6.3. Множественные сравнения.....	445
6.3.1. Поправка по методу Бонферрони.....	447
6.3.2. Поправка по методу Холма	448
6.3.3. Практическое применение поправок по методу Бонферрони и Холма	449
6.3.4. Множественные сравнения и классические статистические тесты	450
6.3.5. Примеры использования критериев для нескольких групп	452
6.4. Коэффициенты корреляции.....	455

6.4.1. Коэффициент корреляции Пирсона	458
6.4.2. Коэффициент корреляции Спирмена.....	461
6.5. Анализ времени до наступления события	462
6.6. Диагностические тесты и порог принятия решения	470
6.6.1. Основные меры точности диагностических тестов.....	470
6.6.2. Использование отношения правдоподобия для расчета вероятности	474
6.6.3. ROC-кривые	476
6.7. Планирование исследования с использованием статистических методов и разведочный анализ данных.....	481
6.7.1. Расчет размера выборки	482
6.7.2. Автоматизация процедуры рандомизации	493
6.7.3. Разведочный анализ данных и концепция воспроизводимых исследований	496

Глава 7. Большие данные и машинное

обучение в медицине	507
7.1. Общее понятие о больших данных.....	507
7.1.1. Признаки и характеристики больших данных	507
7.1.2. Подходы для работы с большими данными	508
7.1.3. Большие данные в здравоохранении	521
7.2. Общие понятия и основы машинного обучения	521
7.2.1. Закодированные правила.....	521
7.2.2. Машинные алгоритмы.....	523
7.2.3. Обучение	528
7.2.4. Переобучение	529
7.2.5. Алгоритмы обучения	536
7.2.6. Искусственный интеллект.....	537
7.3. Математика для машинного обучения.....	537
7.3.1. Введение.....	537
7.3.2. Умножение вектора на скаляр	541
7.3.3. Скалярное произведение векторов.....	542
7.3.4. Сложение и вычитание векторов.....	547
7.3.5. Расстояние между векторами.....	548
7.3.6. Матрица	549
7.3.7. Сложение и вычитание матриц.....	550
7.3.8. Транспонирование матриц.....	551
7.3.9. Умножение матриц.....	551
7.3.10. Определитель матрицы.....	555
7.3.11. Ранг матрицы.....	558
7.3.12. Обратная матрица.....	561
7.3.13. Нормализация	566

7.3.14. Производная	568
7.3.15. Градиентный спуск и функция потерь	573
7.4. Линейная регрессия	576
7.4.1. Условия применимости модели линейной регрессии	577
7.4.2. Обучение модели линейной регрессии	579
7.4.3. Модель линейной регрессии на языке R	584
7.5. Логистическая регрессия	599
7.5.1. Сигмоидная функция	599
7.5.2. Построение логистической регрессии	601
7.5.3. Многоклассовая классификация	610
7.5.4. Построение модели в R	611
7.6. Деревья принятия решений. Случайный лес	617
7.6.1. Деревья принятия решений	617
7.6.2. Случайный лес (random forest)	635
7.7. Бустинг	645
7.7.1. Общие принципы бустинга	645
7.7.2. Градиентный бустинг	651
7.8. Нейронные сети	656
7.8.1. Общие понятия	656
7.8.2. Задачи регрессии и классификации	663
7.8.3. Прямой проход нейронной сети	665
7.8.4. Обратное распространение ошибки	667
7.8.5. Сверточные нейронные сети	674
7.9. Обучение без учителя	703
7.9.1. Алгоритмы кластеризации	703
7.9.2. Реализация алгоритмов кластеризации на языке R	713
7.9.3. Методы понижения размерности	717
7.10. Автоматизация подходов к построению моделей машинного обучения	725

Глава 8. Вспомогательные инструменты для организации

исследовательских проектов	737
8.1. Системы контроля версий	737
8.2. Различные подходы к созданию динамических отчетов	737
8.3. Создание дашбордов с Quarto	738
8.4. Создание интерактивных приложений с Shiny	738
8.5. Организация сбора данных	738

Глава 9. Биомедицинские наборы данных

для машинного обучения	739
9.1. Источники биомедицинских наборов данных	739
9.2. Наборы данных для скачивания	742

Авторы

Кузьменков Алексей Юрьевич — доктор медицинских наук, заместитель директора по стратегическим разработкам НИИ антимикробной химиотерапии, профессор кафедры микробиологии, основатель, руководитель и преподаватель проекта «Цифровая кафедра» с курсом «Технологии науки о данных в медицине» ФГБОУ ВО «Смоленский государственный медицинский университет» Министерства здравоохранения Российской Федерации.

Виноградова Алина Геннадьевна — кандидат медицинских наук, старший научный сотрудник отдела стратегических разработок и проектов НИИ антимикробной химиотерапии, доцент кафедры микробиологии, научный руководитель и преподаватель проекта «Цифровая кафедра» с курсом «Технологии науки о данных в медицине» ФГБОУ ВО «Смоленский государственный медицинский университет» Министерства здравоохранения Российской Федерации.

Трушин Иван Витальевич — младший научный сотрудник отдела стратегических разработок и проектов НИИ антимикробной химиотерапии, преподаватель проекта «Цифровая кафедра» с курсом «Технологии науки о данных в медицине» ФГБОУ ВО «Смоленский государственный медицинский университет» Министерства здравоохранения Российской Федерации.

Авраменко Андрей Алексеевич — младший научный сотрудник отдела стратегических разработок и проектов НИИ антимикробной химиотерапии, технический организатор проекта «Цифровая кафедра» с курсом «Технологии науки о данных в медицине» ФГБОУ ВО «Смоленский государственный медицинский университет» Министерства здравоохранения Российской Федерации.

Кузьменков Михаил Юрьевич — кандидат юридических наук, специалист по машинному обучению отдела стратегических разработок и проектов НИИ антимикробной химиотерапии ФГБОУ ВО «Смоленский государственный медицинский университет» Министерства здравоохранения Российской Федерации.

Предисловие

Наука о данных (Data Science) — это междисциплинарная область, которая объединяет методы, алгоритмы и программные средства, разработанные для извлечения полезной информации из большого объема данных, хранящихся в различных форматах. Соответственно, основной целью Data Science является обнаружение в данных скрытых закономерностей или получение знаний — т.е. такой информации, которая может быть использована для принятия обоснованных и оптимальных решений.

Наука о данных включает в себя различные дисциплины: математическую статистику; системы искусственного интеллекта; управление базами данных; сбор, анализ и визуализацию данных и многие другие. Наибольшее развитие рассматриваемая наука получила с начала 2010-х годов благодаря взрывному росту объемов хранимой в мире информации и вычислительной мощности компьютерной техники. Также немаловажным стимулом к ее развитию стало совершенствование в эти годы алгоритмов машинного обучения, таких как нейронные сети, бустинг, бэггинг и др. На сегодняшний день популярность науки о данных вышла уже далеко за пределы чисто технических сфер, и ее инструменты стали широко применяться в маркетинге, логистике, государственном управлении и — наиболее интересующем нас — здравоохранении.

В медицинской науке, в частности в биомедицинских исследованиях, статистические методы анализа данных использовались еще с середины XX века, что естественным образом создало благоприятную почву для внедрения подходов науки о данных в здравоохранение. К тому же, на сегодняшний день здравоохранение является одной из крупнейших, быстрорастущих и наиболее приоритетных отраслей мировой экономики, которая генерирует огромное количество данных: диагностических, генетических, эпидемиологических, фармакологических, финансово-страховых, медико-социальных и др.

Сегодня технологии Data Science и искусственного интеллекта переосмысливают сферу медицинских услуг, предоставляя уникальные возможности для улучшения диагностики, лечения и профилактики заболеваний. Однако, зачастую, специалисты здравоохранения не обладают достаточными компетенциями ни в сфере информационных технологий, ни в сфере статистического анализа биомедицинских данных для того, чтобы создавать конкурентоспособные продукты. Привлечение в отрасль специалистов по анализу данных (т.н. «дата-сайентистов») с техническим образованием тоже не является панацеей, т.к. для создания эффективных, а главное безопасных медицинских продуктов необходимо глубокое понимание предметной области, связанной со здоровьем человека.

С учетом вышесказанного, наш авторский коллектив, обладающий многолетним опытом преподавания медицинских дисциплин, разработки программных решений для практического здравоохранения, а также создания и реализации курса «Технологии науки о данных в медицине» в рамках проекта «Цифровая кафедра» Смоленского государственного медицинского университета, подготовил для специалистов в области здравоохранения, медицинской информатики и аналитики, а также других заинтересованных специалистов практическое руководство «Наука о данных и искусственный интеллект в медицине». Настоящее издание, делая акцент на практической значимости излагаемого материала, призвано дать читателям комплексное понимание науки о данных в медицине, включающее все важные информационно-технологические и исследовательские аспекты.

Практическое руководство «Наука о данных и искусственный интеллект в медицине» состоит из девяти глав. В первой главе приводятся основные понятия доказательной медицины и принципы организации биомедицинских исследований. Во второй главе рассматриваются основы программирования на языке R, а также алгоритмы сбора и обработки биомедицинских данных. В третьей главе подробно описываются системы управления базами данных, обеспечивающие хранение информации в цифровом структурированном виде. В четвертой главе приводится обзор разнообразных способов компьютерной визуализации, позволяющих наглядно отображать полезную информацию, скрывающуюся в биомедицинских данных. В пятой главе рассматриваются общие, а в шестой — частные вопросы статистического анализа, раскрывающие особенности применения методов математической статистики для формальной оценки и интерпретации биомедицинских данных, а также для формирования обоснованных заключений. В седьмой главе приводятся базовые понятия «Больших данных» (Big Data) и «Машинного обучения» (Machine learning), а также разбираются конкретные примеры реализации алгоритмов машинного обучения для анализа биомедицинских данных. В восьмой главе описываются популярные инструменты и технологии, позволяющие облегчить процесс организации, разработки и представления результатов исследовательских проектов в области науки о данных. Девятая глава дополняет практическую значимость настоящего издания, предлагая подборку реальных биомедицинских данных для отработки навыков, полученных в ходе прочтения.

Мы надеемся, что наш труд будет способствовать притоку в отечественное здравоохранение высококлассных специалистов в области науки о данных и искусственного интеллекта, которые выведут отрасль на новую невиданную высоту.

«Наука о данных и искусственный интеллект в медицине» — уникальное практическое руководство, объединяющее современные подходы анализа больших данных и искусственного интеллекта с прикладными аспектами здравоохранения. Эта книга охватывает весь цикл работы с биомедицинскими данными: от планирования клинических исследований и сбора информации до углубленного статистического анализа и публикации результатов. Читатель познакомится с основами программирования на языке R, различными системами управления базами данных и алгоритмами машинного обучения, а также получит возможность отработать полученные навыки на реальных наборах биомедицинских данных. Издание станет незаменимым помощником как для опытных лечащих врачей и учёных-исследователей, так и для студентов вузов и специалистов из разных областей, желающих овладеть новыми компетенциями.

ISBN 978-5-906023-48-3



9 785906 023483

приоритет

