

Наука о данных: методология, основные направления, проблемы и перспективы

В. И. Городецкий

АО «ЭВРИКА», Санкт-Петербург, Россия

Аннотация. Наука о данных как самостоятельная наука зародилась в начале 2000-х годов как ответ научного сообщества на проблемы обработки больших данных. Сейчас она достигла того уровня зрелости, который делает ее наукой, которая играет важную и незаменимую роль в кластере наук, занимающихся проблемами сбора, хранения, представления и использования данных. В работе очерчиваются границы науки о данных по отношению к искусственному интеллекту. Описываются многоаспектные двухсторонние взаимосвязи науки о данных с другими смежными науками, которые работают с данными. Дается краткое введение в методологию науки о данных, характеризуются основные направления исследований. Перечисляются некоторые источники трудных проблем, решение которых можно ожидать от науки о данных.

Ключевые слова: наука о данных, искусственный интеллект, пирамида DIKW, образ мышления науки о данных, проблемные вопросы науки о данных.

DOI 10.14357/20718594220301

Введение

Мы живем в цифровом мире данных, в котором все, что нас окружает, постоянно измеряется, записывается и сохраняется в цифровой форме. Цифровой мир данных интенсивно развивается, и если пару десятилетий назад этот мир был уделом профессионалов в области информационных технологий (ИТ), то сейчас ситуация поменялась кардинально. Цифровой мир данных стал важной и естественной частью жизни каждого человека и его деятельности. Причины, которые обусловили такую «экспансию» цифрового мира данных и которые способствуют дальнейшему расширению этой экспансии, хорошо известны, и снова обсуждать их здесь вряд ли стоит.

Относительно недавно, где-то 10 – 15 лет тому назад, когда в научном обиходе появился термин «большие данные», специалистам стало

понятно, что они являются чрезвычайно ценным ресурсом. Однако попытки использовать этот ресурс в соответствии с его ценностью, да и просто попытки понять истинную ценность тех или иных данных оказались не столь продуктивными, как на это надеялись. Вместе с большими данными пришли и большие проблемы, и вызовы. Оказалось, что мы знаем о данных совсем мало, что данные изучены плохо и что они являются специальным объектом научных исследований.

Каждый набор данных содержит в себе скрытые знания, которых нет в другом наборе, но мы пока далеко не всегда знаем, как подступить к поиску этих знаний. В работе [1] отмечается, что мы все еще по-прежнему мало знаем о методах и технологиях работы с данными разных типов, объемов, размерности и сложности. Мы пока плохо знаем, что такое потенциал, скрытый в данных, и не умеем в пол-

✉ Городецкий Владимир Иванович. E-mail: Vladimir.gorodetsky@gmail.com

ной мере определять и извлекать ценность, содержащуюся в них, не умеем эффективно конвертировать данные в нечто другое, приносящее если не прибыль, то пользу.

Наука о данных стала ответом на те проблемы и вызовы, с которыми встретились исследователи и разработчики в области ИТ при работе с большими данными. Эта наука имеет целью найти решения проблем, некоторые из которых названы выше, а также ответить на многочисленные новые вызовы, которые появляются по мере изучения данных и роста зрелости науки о них.

Объектом исследований этой новой науки являются данные, их свойства, сложности, разнообразные связи в данных и их практическое использование. Эта наука изучает знания и интеллект, скрытые в данных, методы и технологии преобразования данных от сырых сигналов к информации и знаниям [1 – 7]. Она изучает методы, модели и алгоритмы преобразовании знаний, полученных из данных, к управленческим решениям и к действиям во исполнение этих решений, а также исследует и предлагает модели данных, подходящие для решения тех или иных прикладных задач.

Данная работа имеет целью очертить место науки о данных в контексте современных наук кибернетического кластера, научные и прикладные результаты которых используются в основе науки о данных, и подчеркнуть ее отличие от искусственного интеллекта (ИИ) и других смежных наук, дать краткое введение в методологию науки о данных как самостоятельной науки и очертить ее базовые научные направления. Кроме того, в работе перечисляются некоторые трудные проблемы, решения которых ожидают от науки о данных.

1. Данные и знания

Наука о данных занимается систематическим изучением данных и их свойств с разных точек зрения, методологическим и алгоритмическим обеспечением их полного жизненного цикла. Этот цикл обычно включает в себя сбор данных, их очистку, несколько этапов преобразования к удобной форме, диктуемой целевой задачей, и практическое использование данных для решения одной или нескольких задач предметной области, к которой данные относятся. Кроме того, жизненный цикл данных включает в себя хранение данных, что не сводится к про-

стым процедурам записи данных на физический носитель. Хранение данных требует использования некоторых технологий, которые должны обеспечить заданные показатели качества процессов хранения и доступа к данным, а также требуемый уровень доверия к ним и безопасности. Исследование таких аспектов хранения данных, создание алгоритмических и инструментальных средств, гарантирующих нужный уровень качества хранения данных, – это тоже область исследований и разработок науки о данных.

Однако столь широкие притязания науки о данных вызывают естественный вопрос о том, а что же тогда относится к другим наукам, которые тоже работают с данными, например, к статистике, вычислительным и когнитивным наукам, к ИИ? Можно ли как-то убедительно очертить границы науки о данных в контексте ее ближайших соседей по научному полю? В конце концов, каждая предметная область тоже работает со своими данными и зачастую использует свои специфические подходы и научные основы. Например, бизнес-аналитика немислима без обработки данных, как и многие другие теоретические и прикладные науки, и использует данные в своих интересах.

Вопрос о границах науки о данных достаточно активно обсуждался и продолжает обсуждаться научной общественностью. Скорее всего, этот вопрос будет обсуждаться еще долго. Чтобы в нем как-то обоснованно разобраться, нужно сначала проанализировать различные информационные категории, которыми оперируют специалисты в области ИТ при работе со структурами данных разной сложности, называя их информацией, знаниями и др.

Предмет исследования науки о данных в ИИ – это данные [1-7], предметом исследований являются знания [8]. Другие науки тоже вносят свой вклад в исследования, связанные с данными и знаниями. Поэтому прежде чем обсуждать вопрос о соотношении науки о данных и других наук, близких к ней, наверное, логично сначала как-то очертить границы, определить сходство и различия между данными, знаниями, информацией и другими информационными структурами, и если этот вопрос не решается однозначно, то просто уточнить смысл этих понятий в рамках материала, излагаемого далее.

Обсуждение существа понятий *данные, информация, знания* и других аналогичных поня-

тий в контексте Информационной науки¹ было впервые начато в [9]. Ее автор, известный специалист в области теории систем Р. Аккофф (R.L. Ascoff), ввел понятие *когнитивной иерархии* на множестве понятий *данные, информация, знания, интеллект и мудрость* (англ. *data, information, knowledge, intelligence, wisdom, DIKIW*). При этом под когнитивным уровнем автор понимал сравнительную характеристику глубины человеческого познания, которая нашла отражение в структурах данных, представляющих перечисленные понятия.

Эта иерархия, в настоящее время широко известная под названием *Пирамида DIKIW*, представлена на Рис. 1 [10]. Свое понимание смысла понятий, упорядоченных в этой пирамиде, Р. Аккофф описал так [9]:

– «*Данные* определяются как символы, которые представляют свойства объектов, событий и их окружения. Они бесполезны до тех пор, пока не будут представлены в форме, пригодной для использования. ... Данные являются продуктом наблюдений.

– *Информация* содержится в описаниях и в ответах на вопросы к данным, которые начинаются с таких слов, как *кто, что, когда и сколько*. ... Информация выводится из данных.

– *Знания* – это ноу-хау, это то, что делает возможным преобразование информации в инструкции. Знания могут быть получены либо путем передачи от другого человека, который ими обладает, либо путем инструкции, либо путем извлечения из опыта.

– *Интеллект* – это способность повышать эффективность знаний.

– *Мудрость* – это знание *почему?*, способность повышать эффективность знаний и добавлять к ним ценность, а это требует наличия когнитивной функции, которую мы называем суждением.»

На заре зарождения современной науки о данных вопрос о смысле понятий пирамиды DIKIW был темой международного проекта США “Knowledge Map of Information Science” (Карта знаний информационной науки), который выполнялся в 2003 – 2005 гг. с целью изучения оснований и строительных блоков этой



Рис. 1. Пирамида преобразований DIKIW: «данные, информация, знания, интеллект, мудрость»

науки. Например, в работе [11] (это одна из четырех работ, в которых подводились итоги проекта) дано 130 определений понятий *данные, информация и знания*, отражающих основные подходы к их определению. Они сформулированы 45 учеными из 16 стран. Основной вывод этого проекта состоял в том, что эти три ключевых понятия взаимосвязаны, но характер отношений между ними является дискуссионным, равно как и их смысл.

Глубокий анализ этой же проблемы представил автор [12]. В итоге он пришел к выводам, аналогичным выводам работы [11]: «...остаются спорными вопросы к тому, как выполняются преобразования понятий пирамиды DIKIW, находящихся в иерархии ниже, в понятия, находящиеся выше». Именно это, по мнению автора [12], приводит к отсутствию ясности в их определении.

Если рассматривать понятия этой пирамиды не в контексте когнитивного уровня, который представлен в них, а с позиций сложности структур данных, которые требуется использовать для их формального описания (представления), то такую тоже качественную характеристику данных называют *информационной зрелостью* понятий, или, по-другому, *уровнем знаний* (интеллекта), представленного в модели данных [1], которая может быть, в частности, компьютерной программой. Заметим, что отличие когнитивного уровня модели от уровня знаний, представленных в модели средствами ИИ, описано в [13].

Качественная иллюстрация соотношения между уровнем когнитивной зрелости понятий пирамиды DIKIW, уровнем интеллекта, представленного в них, и базовыми понятиями *данные, информация, знания, интеллект и муд-*

¹ Деление наук за рубежом отличается от принятого в России. Содержание науки “Information Science” см. по ссылке [https://en.wikipedia.org/wiki/Information_science#:~:text=Information%20science%20\(also%20known%20as,dissemination%2C%20and%20protection%20of%20information.](https://en.wikipedia.org/wiki/Information_science#:~:text=Information%20science%20(also%20known%20as,dissemination%2C%20and%20protection%20of%20information.)

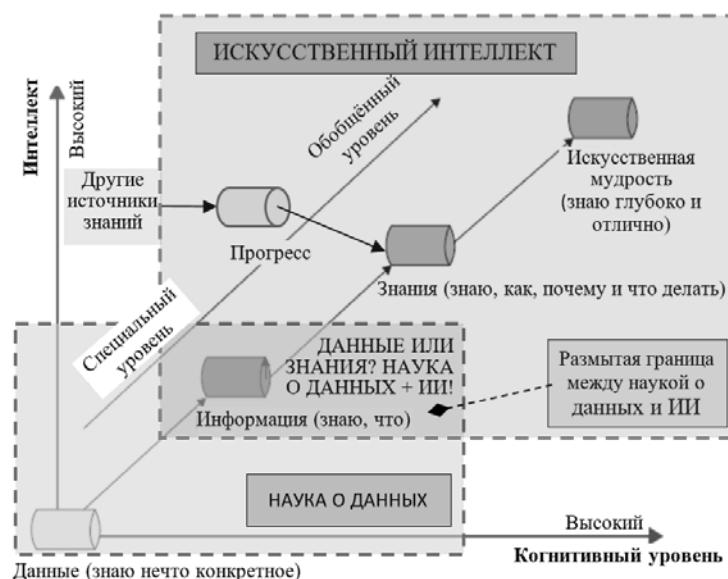


Рис. 2. Когнитивный прогресс процесса трансформации
данные → информация → знания → мудрость

рость² иллюстрируется на Рис. 2, заимствованном из [1]. На этом рисунке иерархия информационных структур пирамиды DIKIW представлена точками на биссектрисе угла между координатными осями «когнитивный уровень модели данных» (по горизонтали) и «уровень интеллекта, представленного в модели данных» (по вертикали). Полагается, что уровень «информационной зрелости» понятий, расположенных вдоль этой биссектрисы, возрастает от «слева – снизу» в направлении «вправо – вверх». По замыслу авторов пирамиды, и этому следует автор [1], она должна качественно проиллюстрировать разницу между понятиями *данные*, *информация*, *знания*, *интеллект*, и *мудрость* тем, что каждый последующий уровень пирамиды DIKIW обладает более высоким когнитивным уровнем и интеллектом, встроенным в модель соответствующей информационной структуры [1]. Дадим современное толкование смысла понятий пирамиды DIKIW, опираясь на мнение и описания автора одной из недавних работ по методологии науки о данных [1], и имея целью пояснить разницу между этими понятиями и наметить их границы.

Данные представляют дискретные или непрерывные *факты*. Они могут быть составлены из символов, могут быть сигналами сенсоров, измеряющих некоторые свойства физических, или виртуальных объектов или субъектов. Дан-

ные лежат на нижнем уровне когнитивной системы, они могут быть субъективными или объективными, могут иметь или не иметь смысл и могут иметь определенную ценность. Примерами данных могут быть бинарная последовательность 110, число 3,1415, слово «молодой», и т.п.

Информация есть описание данных, конкретных объектов или субъектов любого типа (физических, виртуальных, социальных) в некотором организованном виде, которое для потребителя данных имеет конкретный смысл. Информация – это данные, к которым добавлен *смысловой контекст*. Например, «110 – двоичное представление числа 7», «Число π равно 3,1415», «Петр молодой ученый», и т.п. Информация уже имеет определенную цель и/или смысл. Она может быть структурированной или функциональной, субъективной (качественной оценкой человека, например, или относящейся к его намерениям) или объективной (основанной на фактах, измерениях). Например, «Петр молодой ученый» — это субъективная информация, представляющая отношение между двумя сущностями: «Петр» и «молодой ученый», а информация «Пуд равен 16 кг» есть объективная информация, представляющая или результат взвешивания пудовой гири, или результат договоренности о том, что такое пуд веса. Фактически *информация* задается отношениями на множестве данных и сущностей в контексте, определяющем смысл данных.

²Понятие *интеллект* в этой пирамиде опущено для того, чтобы не перегружать рисунок.

Знание представляет собой структурированную форму информации, например, структурированную смесь информации, некоторые правила, заданные описанием процедуры их реализации, или их декларативным представлением, например, в форме продукций. Знание, как и данные, может быть субъективным или объективным, относиться к физическим и/или виртуальным или социальным объектам и субъектам. Знание может быть истинным или ложным. Оно может быть убеждением субъекта, но в реальности может быть ложным. Оно может быть известным или неизвестным, применимым на практике и неприменимым, разумным или нет. Примерами знаний, описанных процедурно или представленных пропозициональными правилами, могут быть, например, такие, как «Молодые ученые имеют возраст до 40 лет», «Чтобы стать студентом, нужно сдать вступительные экзамены», «Если человеку больше 65 лет, то он может выйти на пенсию». Знание *добавляет* к информации механизм ее использования (как?), объяснение (почему?) и/или целеполагание (что делать?). Формально знания, как и все другие информационные структуры пирамиды DIKW, кроме данных, задаются отношениями на множестве сущностей ее нижележащих уровней.

Интеллект (на Рис. 2 эта позиция отсутствует) отвечает понятиям пирамиды, которые можно использовать для обобщения информации или знаний, их исследования, для выполнения рассуждений об информации и о знаниях, получения выводов на их основе или путем обработки их с целью получения новых знаний. Интеллект может быть врожденным, может быть приобретён путем обучения или личного опыта. Интеллект может иметь свои градации и потому иногда возможно сравнение разных уровней интеллекта. Уровни интеллекта могут быть упорядочены иерархически, например, интеллект может быть более высоким и менее высоким, быть более общим или специфическим. Примеры интеллектуальных высказываний: «Возраст Петра меньше 40 лет» (результат вывода, основанного на фактах «Петр является молодым ученым» и «Молодые ученые имеют возраст до 40 лет»).

Мудрость представляет собой информационный ресурс, который, кроме компонент интеллекта, знаний и информации, содержит еще высокоуровневые правила, которые являются

результатом их обобщения на метауровне. Высокоуровневые правила могут также быть результатом индуктивного вывода или другого процесса обработки знаний и интеллекта творческого характера с использованием метазнаний, интуиции в тех случаях, когда носителем мудрости является человек, принимающий участие в процессах работы с информационными структурами. Он может привносить в эти процессы свою интуицию и вдохновение, приобретенные опытным путем в процессе интеллектуальной деятельности. Мудрость может быть нематериальной, уникальной, личной, интуитивной или быть результатом умственного вдохновения. По сравнению со знанием мудрость находится вне времени. Она является всесторонней, общей и эмоциональной. Мудрость передается из поколения в поколение и между культурами в форме поговорок, цитат и философских высказываний.

Заключая описание информационных структур, упорядоченных в пирамиде DIKW, можно сделать вывод о том, что данные, информация, знания, интеллект и мудрость связаны отношением обобщения, отношением частное – общее, а последовательный переход от данных к мудрости – это переход, который характеризуется увеличением когнитивного уровня модели данных и ее интеллекта.

Рис. 2 позволяет очертить границы науки о данных и понять соотношение ее научной области и соответствующей области ИИ. Наука о данных решает, в основном, задачи, связанные с данными и информацией, и это основная зона ее научных и прикладных интересов. На рисунке это отражено прямоугольником, помеченным названием «Наука о данных».

Предметом исследований ИИ являются знания. Но информационные структуры типа *знания* и *информация* точно отделить друг от друга невозможно, поскольку понятие *информация*, как и понятие *знания*, формально описываются отношениями на множестве тех или иных сущностей, хотя для информации характерны более простые информационные структуры. Но это качественное различие. По этой причине в область интересов науки о данных могут попасть информационные структуры типа *информация*, которую будет трудно отличить от знаний.

На Рис. 2 область интересов ИИ представлена прямоугольником, который содержит внутри себя все информационные структуры, кроме

данных. Поэтому структуры, которые названы на этом рисунке *информацией*, являются областью пересечения интересов ИИ и науки о данных. Это полностью совпадает с выводами проекта США «Карта знаний информационной науки», о котором шла речь раньше [11, 12]. Соответствующие научные проблемы науки о данных следует отнести к ее междисциплинарному контексту с ИИ.

2. Место науки о данных в контексте смежных наук

Исследование и использование данных – это не новая задача. Ее возраст насчитывает уже несколько столетий, если не больше. По этой причине, когда примерно в 2010 г. специалисты в области обработки данных стали говорить о рождении новой науки, науки о данных, то это вызвало много вопросов, и многие участники дискуссий не хотели соглашаться с тем, что науку о данных следует рассматривать как новую самостоятельную науку. Заметим, что такие дискуссии не закончились и к настоящему времени.

Основными оппонентами новой науки были специалисты в области статистики, которые рассматривали новые методологии, методы и алгоритмы, используемые в науке о данных, просто как развитие статистики. И их позиция была достаточно убедительной, поскольку статистические модели и методы оставались основными при решении задач науки о данных. Заметим, что до сих пор статистика составляет важную компоненту основ науки о данных.

Однако новая наука довольно быстро завоевала право на самостоятельность. Этому сильно способствовало осознание проблем больших данных (примерно 2009 г.), для работы с которыми традиционные методы аналитики и статистики оказались не готовыми. Для того, чтобы справиться с новыми проблемами и вызовами, обусловленными большими данными, потребовалось пересмотреть методологические, методические, алгоритмические и инструментальные средства работы с данными, которые на тот момент использовались.

Оказалось, что классическая статистика вычислительно неэффективна при работе с большими данными и работает с ними неустойчиво, а потому нуждается в новых подходах и моделях, способных справиться с обработкой гете-

рогенных данных больших объемов и размерности и с решением задач статистического машинного обучения на их основе [14, 15]. Такие подходы и модели были созданы в науке о данных [16 – 18].

Кроме того, для больших данных потребовались новые подходы и методы, которые не относятся к статистике. Это проблемы инжиниринга признаков, редукция пространств данных, работа с разреженными матрицами большой размерности, обеспечение робастности вычислений, интеграция гетерогенных данных из многих источников и приложений. Проблемы обработки больших данных потребовали пересмотра инфраструктурных решений для поддержки распределенных и параллельных вычислений, которые стали необходимы для решения проблемы вычислительной эффективности обработки больших данных. Соответствующие исследования активно велись в области науки о данных, расширяя ее методический и алгоритмический базис, а также вычислительную инфраструктуру и другие компоненты экосистемы обработки и анализа больших данных. Эти исследования привели к тому, что статистический базис науки о данных, хотя он и сохранил за собой важную роль в ней, перестал быть монополистом, поскольку фундаментальные основы науки о данных получили значительное развитие за счет других наук.

В результате наука о данных стала непохожа ни на одну из тех наук, методологии, методы и алгоритмы которых были интегрированы в ней. Более того, наука о данных стала активным «экспортером», собственных методов и инструментов в другие науки. В результате этих процессов наука о данных, изначально междисциплинарная наука, которая использует достижения, методы и модели различных наук, стала и трансдисциплинарной наукой, методы которой стали использоваться и трансформироваться в большом числе различных наук, в том числе и в таких, в которых ранее об использовании численных моделей и методов работы с данными даже не возникало мысли. Примером здесь являются все общественные науки типа истории, социологии, политологии и многие другие науки гуманитарной направленности.

Наиболее глубоко междисциплинарные и трансдисциплинарные связи и корни науки о данных проанализированы в [1]. Дадим по этому поводу краткие сведения, опираясь на со-

держание этого источника. Кроме чисто исторического контекста науки о данных, этот вопрос представляется важным и с точки зрения образования в области науки о данных. Ответ на вопрос о корнях науки о данных, ее содержании и связях с другими науками позволит лучше ориентироваться в множестве возможных сочетаний различных университетских курсов в зависимости от специальности, по которой ведется обучение специалистов по науке о данных. Заметим, что вопрос об образовательных программах подготовки по разным специальностям в области науки о данных детально освещён в [1] и в другой литературе открытого доступа.

К числу наук, которые имеют заметное пересечение с наукой о данных, обычно относят статистику и математику, когнитивистику (когнитивную науку, *англ.* cognitive science), информационные науки, включая информатику и искусственный интеллект, науки об интеллекте (*англ.* intelligence science), вычислительную технику, общественные науки, менеджмент и ряд других [1]. Эти науки играют разную роль в задачах науки о данных, и автор [1] делит их на две группы.

В первую группу, по его мнению, входят науки, которые формируют *ядро науки о данных* и которые не зависят от предметной ориентации приложений науки о данных и их особенностей. К этой группе отнесены статистика и математика, когнитивные науки, информационные науки, наука об интеллекте и вычислительная техника. Вторая группа объединяет науки, которые не являются базовыми и методы которых используются при необходимости, обычно привлекаются при решении специальных вопросов (в [1] они называются «мягкими» основами науки о данных). К ним относятся общественные науки, менеджмент, коммуникационные науки и некоторые другие. Рис. 3 иллюстрирует этот список основ науки о данных графически. Дадим краткие сведения о взаимодействиях науки о данных и наук, входящих в ядро науки о данных и в ее «мягкие основы» [1].

Статистика и математика по-прежнему играют ключевую роль в науке о данных. Заметим, что сам термин «наука о данных» впервые был предложен специалистами по статистике. Сейчас связи между статистикой и математикой, с одной стороны, и наукой о данных, с другой, носят двухсторонний характер. Стати-

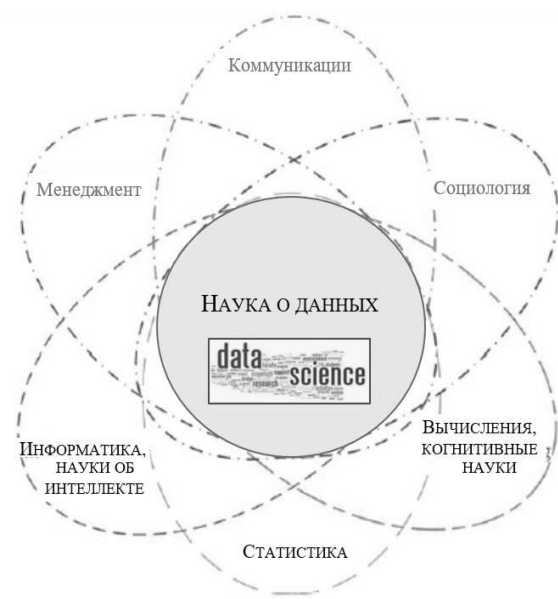


Рис. 3. Междисциплинарный контекст науки о данных

стика предоставляет в распоряжение исследователей данных различные и разнообразные теоретически корректные методы количественного характера и инструменты анализа данных. Примерами являются статистические методы описания данных, методы статистического вывода и предсказания данных, а также различные статистики для описания свойств выборок данных. Наука о данных использует методы корректного формирования выборок данных (*англ.* sampling) и строгое обоснование понятия эмпирических вероятностей и методов работы с ними. Статистика предоставляет понятие и методы квантификации корреляционных зависимостей в данных, оценки их точности и алгоритмы практического использования при решении одной из главных задач науки о данных, которую принято называть анализом связей и поиском отношений в данных (*англ.* Data Mining). Статистика и специальные математические методы составляют основу при анализе пространств данных и, наряду с другими науками, она играет важную роль в инжиниринге признаков и пространств признаков [1].

Но традиционная статистика нуждается в серьезном расширении для того, чтобы работать с крупномасштабными данными, гетерогенными, темпоральными и высокочастотными данными, с данными, связанными многими отношениями и данными с неопределенностью. Статистика нуждается в методах устойчивой работы с разреженными данными и с данными,

которые имеют пропущенные значения. И все перечисленные методы должны быть вычислительно эффективными и устойчивыми. А эти проблемы решают модели и методы, в том числе статистические, созданные в науке о данных, которые стали и инструментом статистики благодаря науке о данных [1, 2, 16 – 18].

Когнитивистика – это междисциплинарная наука, которая находится на стыке теории познания, когнитивной психологии, нейробиологии, философии, антропологии, физиологии, лингвистики и искусственного интеллекта [1]. Ее основным предметом исследований являются процессы познания человека и их представление в его мозге. В науке о данных используются модели когнитивистики, которые относятся, главным образом, к механизмам и методам человекоподобных вычислений и рассуждений, к механизмам обучения и алгоритмам принятия решений и распознавания в сложных многошаговых задачах. В науке о данных привлекаются также результаты когнитивных исследований механизмов памяти, модели восприятия, воображения и представления данных и информации для создания человекоподобных алгоритмов работы с ними, в частности, алгоритмов поиска связей и зависимостей в данных и машинного обучения на основе сложных данных для решения трудных проблем. Методы когнитивистики способствовали развитию нейроподобных структур обработки данных для формирования систем принятия решений и обучения с подкреплением. Методы науки о данных, в свою очередь, используются в когнитивных науках для создания моделей в задачах вычислительного и образного понимания данных и моделей мышления. Оно может стимулировать создание новых архитектур компьютеров и вычислительных систем с новыми способами хранения и обработки больших данных, их грануляции и иерархического представления [1].

Информационная наука и информатика. В зарубежной литературе информационной наукой называют междисциплинарную область, в которую включают разделы информатики, изучающие проблемы доступа к информации, ее сбора, хранения, анализа, защиты и публикации. К информатике относят проблемы обработки информации и создания информационных систем или, говоря другими словами, информатика «представляет теории, средства и системы для преобразования данных в информацию в целях ее обработки и управления ею» [1, с. 169].

Взаимодействие информационных наук и информатики, с одной стороны, и науки о данных с другой, носит столь тесный и многообразный характер, что до сих пор некоторые специалисты отдают в этом взаимодействии ведущую роль науке о данных, а другие уже полагают, что в этой паре ведущая роль все же принадлежит информационным наукам и информатике. Действительно, наука о данных активно использует результаты информационной науки и информатики. Не случайно они отнесены к ядру науки о данных. Если говорить кратко о роли информационных наук и информатики в науке о данных, то нужно отметить, что она использует их результаты в задачах управления данными при интеграции данных и приложений, в задачах представления данных при их хранении и обработке, например, в интересах аналитики данных использует информационные системы, сети и инфраструктуры, созданные на основе информационных технологий [1].

Однако после 2010 г. ландшафт данных сильно изменился, изменились и их свойства. Например, появились данные социальных медиа огромного объема, в которых значительный вес занимают аудио- и видео материалы и тексты на естественном языке (ЕЯ), многоязычные тексты, а также синтетические мультимедиа-данные. Значительный удельный вес в этот период приобрели проблемы семантического анализа медиа- и бизнес-данных, сформировалось научное направление в работе с данными, которое названо «семантическими вычислениями». Естественно, что к работе с такими данными и с семантикой данных традиционная информационная наука и информатика оказались просто не готовы. Эти задачи становятся возможным решать благодаря новым разработкам науки о данных, а потребителями этих разработок стали и задачи информационных наук, и информатики, и науки о данных с приложениями от бизнеса, и финансового анализа до литературного и политического анализа.

Подробный анализ взаимодействия науки о данных и информационных наук совместно с информатикой можно найти в [1, гл. 6].

Наука об интеллекте. В зарубежной литературе наукой об интеллекте называют область исследований, которая охватывает такие разделы как вычислительный интеллект, распознавание образов, обнаружение связей и зависимостей в данных, машинное обучение, компьютерное зрение и различные виды аналитики данных, например,

мультимедийная аналитика [1]. Сюда же принято относить и методы вычислительного интеллекта, инспирированные природой, и эволюционные вычисления.

Модели, методы, алгоритмы и инструменты этих разделов науки об интеллекте имеют уже более чем пятидесятилетнюю историю, и до появления больших данных исследования в этих областях ориентировались, в основном, на работу с наблюдениями и транзакциями, представленными в таблицах или базах данных. В отличие от этого, наука о данных фокусируется на поиске скрытых закономерностей, связей, аномалий и трендов в данных сложной структуры, большого объема и размерности типа текстов на ЕЯ, аудио- и видеоданных, мультимедиа-данных, которые в области больших данных составляют более 90% всех данных. Основные методы обработки таких данных были созданы исследователями в области науки о данных. Соответствующий раздел науки о данных является существенным расширением науки об интеллекте, хотя и активно использует его методы, алгоритмы и инструменты, созданные в период до середины 2000-х годов. В число достижений науки о данных в этой области входят методы, алгоритмы и инструменты для решения задач, которые в науке об интеллекте имеют те же самые названия, а именно обнаружение знаний, скрытых в данных, поиск зависимостей и связей в данных, обучение на основе модели, машинное обучение, компьютерное зрение, распознавание образов, предсказательная и глубинная аналитика, вычислительный интеллект.

Но существенное отличие перечисленных методов в науке о данных от методов науки об интеллекте с аналогичными названиями состоит в том, что наука о данных работает с данными, которые по сложности, разнообразию, размерности и объему не сравнимы с данными, на работу с которыми способны методы науки об интеллекте.

Наука о данных – это следующее поколение информационной науки, информатики и науки об интеллекте. Она имеет целью решение задач на самом низком уровне пирамиды DIKW (Рис. 2), и этим она отличается от искусственного интеллекта, который работает и со знаниями, и с интеллектом, и с мудростью.

Вычислительная техника. Связь науки о данных с вычислительной техникой обусловлена тем, что вычисления, наряду с другими науками, формируют базис науки о данных, ее фундаментальную компоненту. Вычисления ре-

ализуют алгоритмы и другие процессы науки о данных, и она формирует к ним новые жесткие требования, стимулируя тем самым развитие вычислительной техники. Эти требования формулируются к ее инфраструктуре, архитектуре, памяти, логистике данных и программной платформе. Они направлены на ускорение вычислений за счет параллельных, распределенных и иерархически организованных вычислений, за счет облачных, мобильных, периферийных и встроенных вычислений. Требования со стороны науки о данных касаются также повышения скорости передачи сообщений и обеспечения гибкости адресации, повышения пропускной способности коммуникационных каналов, а также обеспечения гибкого автономного управления процессами вычислений и коммуникаций. Например, традиционные архитектуры вычислительных систем обладают неудовлетворительными характеристиками по производительности. Суперкомпьютеры в какой-то мере ослабляют эту проблему, но они работают с требуемой скоростью с данными, которые полностью размещаются в оперативной памяти. Суперкомпьютеры обладают «плохой логистикой» данных: если возникает необходимость обмена данными с внешней памятью, то скорость вычислений может упасть на несколько порядков. Взаимное влияние науки о данных и вычислительной техники в этой задаче состоит в том, что наука о данных стимулирует разработку новых вычислительных архитектур и новой элементной базы, например, программируемых логических интегральных схем (*англ.* FPGA, Field-Programmable Gate Array), а в науке о данных разрабатываются специальные алгоритмы, которые не требуют обращения к внешним хранилищам данных или минимизируют число таких обращений.

Что касается наук, которые формируют «мягкий» базис науки о данных, к которым отнесены общественные науки, менеджмент и коммуникации между людьми, то их взаимодействие с наукой о данных также имеет двухсторонний характер, хотя, бесспорно, влияние науки о данных на них несравненно более сильное, чем влияние в обратном направлении. По существу, численные методы и модели и специализированные варианты анализа данных, относящихся к наукам «мягкого» базиса стали возможными благодаря проникновению методов и моделей науки о данных в эти области

научного знания. Можно сказать, что методы науки о данных привели к революционным преобразованиям наук гуманитарной сферы.

Но и вклад в обратном направлении, в частности, вклад общественных наук в науку о данных тоже существенен [1]. В ней социальные методологии исследований нашли применение при решении различных задач работы с данными и принятия решений на их основе. В частности, методы ведения переговоров, формирование рабочих групп и коалиций для группового принятия решений, методы поиска консенсуса, качественные методы исследования, качественные модели поведения и методы их исследований, поведение организаций, управление отношениями и коммуникациями в коллективах, модели лидерства и коллективной оценки решений – это то, что пришло в науку о данных со стороны наук «мягкого» базиса.

3. Методологические основы науки о данных

Хорошо известно, что любая наука, кроме своего специфического объекта исследования, каким для науки о данных являются данные, должна иметь и свою *методологию* исследований, которая отличает ее от других наук. Методология науки о данных, которую иногда называют «образом мышления в науке о данных», имеет ряд специфических особенностей, которые действительно сильно отличают ее от других наук, работающих с данными.

Методология науки о данных в обобщенном виде «разделяет действия по решению задач на четыре уровня, реализуемых последовательно [1, гл. 3, с. 72]:

- 1) исходный уровень — уровень исходных данных;
- 2) уровень выбора механизмов решения задач науки о данных;
- 3) уровень предоставления результатов заказчику;
- 4) уровень гарантий качества результатов.»

Для каждого из этих уровней методологические принципы науки о данных формулируются следующим образом [1]:

– Данные являются основным исходным материалом для решения задач; эти решения строятся на основе данных.

– Изучение и исследование данных проводится без предварительного выдвижения гипотез.

– Обнаружение знаний в данных не опирается на предположения о модели данных.

– Принятие решений базируется на фактах и эмпирических доказательствах.

Поясним существование этих методологических принципов.

Решения строятся на основе данных. В науке о данных полагается, что данные — это основной вход научного исследования. Естественно, что всегда присутствуют и другие источники, например, экспертные знания о приложении, известные результаты предметной области. Но в любом варианте данные полагаются основным входом, на который следует полагаться при решении задач науки о данных.

При выполнении исследования его ход не является заранее известным или предсказуемым. Стратегия исследований должна определяться данными, именно данные должны диктовать, что делать на очередном этапе. Например, процесс исследований, управляемый данными, определяет используемые методы, алгоритмы и модели.

«Идеальное исследование — это процесс, который в совершенстве подгоняет друг к другу данные наблюдений, обучаемую модель и результаты, получаемые онлайн в ходе исследований. Это совсем не то, что процесс подгонки данных к *заданной* модели или подгонки *заданной* модели к данным.» [1, гл. 3, с. 81]

Исследование проводится без выдвижения гипотез. Традиционное научное исследование, в том числе, и связанное с данными, состоит в том, что сначала выдвигается некоторая гипотеза, например о независимости и одинаковом распределении атрибутов данных, или о нормальном распределении данных. Далее собираются или генерируются экспериментальные данные, которые используются для того, чтобы доказать и/или опровергнуть выдвинутую гипотезу. Но иногда о проверке гипотез просто не заботятся. Подчеркнем, что таков основной путь традиционного исследования в разных науках, и в них данные используются для того, чтобы проверить истинность выдвигаемых гипотез.

Наука о данных использует иную методологию, в которой никакие гипотезы о свойствах данных априори не выдвигаются и во внимание не принимаются. Полагается, что источником гипотез могут быть только сами данные, а также дополнительно априори известные свойства

предметной области, к которой относятся данные. Для науки о данных данные первичны, и гипотезы об их модели должны строиться по результатам исследования данных. Исследователь в области науки о данных имеет изначально множество данных и использует их для установления закономерностей, зависимостей и связей.

Данные обычно относятся к той или иной предметной области, для которой могут быть достоверно известны какие-то закономерности, отношения или связи. Последние не являются гипотезами. Например, данные, которые относятся к траектории баллистической ракеты могут быть результатами реальных измерений, но в научных исследованиях эти данные обычно получаются путем интегрирования системы дифференциальных уравнений движения, известных из баллистики. В реальности любая реализация такой траектории подвержена непредсказуемым воздействиям как со стороны внешней среды (ветер, вариации атмосферного давления и плотности, возмущения поля земного тяготения и т.п.), так и со стороны системы управления (например, вследствие отклонения модуля вектора тяги двигателя и его пространственной ориентации от их номинальных значений и др.). С помощью генерации выборки случайных траекторий движения ракеты на основе ее модели можно построить некоторое множество экземпляров траекторий. Все экземпляры этого множества, а именно случайные реализации траекторий, будут принадлежать некоторому многообразию в пространстве переменных движений и атрибутов уравнений, причем для всех элементов этого многообразия будут выполнены предметные закономерности, автоматически учтенные в модели движения ракеты. Предположим, что требуется построить модель отклонений координат и скоростей траектории ракеты в конечной точке траектории от расчетных значений в зависимости от тех или иных случайных возмущений. Например, пусть решается задача выявления наиболее сильно влияющих возмущений, чтобы затем использовать полученный результат для повышения точности управления программным движением. Если такая задача решается с использованием методологии науки о данных, то никаких гипотез или предположений о модели этих зависимостей не выдвигается. Имеющиеся данные дополняются данными моделирования, и вся

полученная совокупность данных используется для исследования и решения поставленной задачи. Гипотезы (модели искомых зависимостей) должны быть установлены на основе анализа данных.

Обнаружение знаний в данных не опирается на предположения о модели данных. Процесс обнаружения знаний в данных основывается только на данных, в отличие, например, от методологии обнаружения знаний на основе модели. В науке о данных эта модель неизвестна, и ее нужно построить.

Специалист в области науки о данных на начальном этапе построения модели, например, модели принятия решений, имеет только данные и решает задачу обнаружения знаний в данных, нужных для принятия «хороших» решений, опираясь только на эти данные.

Одна из основных задач, которую он должен решить при анализе данных, – это задача инжиниринга пространства атрибутов будущей модели, или, как часто говорят, задача выбора наиболее информативных атрибутов целевой модели. В отличие от этого, в традиционном подходе структура модели и ее атрибуты обычно задаются, а данные используются для оптимизации значений выбранного множества атрибутов. Как образно замечено в [1], в методологии науки о данных «данные должны сами рассказать о себе». Другими словами, построение модели для решения прикладной задачи в науке о данных рассматривается как процесс, управляемый самими данными.

Естественно, что это сложный процесс, и он не всегда в итоге приводит к построению наилучшей модели, например, наилучшей модели принятия решений. Но наука о данных располагает различными методами оценки качества построенной модели и проверки ее практической пригодности (*англ.* validation). Если данные относятся к предметной области, для которой известны определённые свойства и закономерности, то их необходимо учитывать при работе с данными. Но эти знания не являются гипотезами о данных. Пример, приведенный выше для баллистической ракеты, иллюстрирует роль модели предметной области в методологии науки о данных.

Отметим, что на практике методология исследований в области науки о данных иногда комбинируется с другими методологиями, например, с экспериментальными исследовани-

ями и имитационным моделированием. Однако еще раз подчеркнем, что в науке о данных обнаружение знаний основывается на анализе самих данных, хотя на каких-то этапах этого анализа процесс машинного обучения может основываться и на модели [1]. Примером является использование модели линейной регрессии после инжиниринга признаков и визуального изучения данных. Другой пример – это предположение о компактности множеств примеров, которые относятся к одному и тому же классу, принимаемое в некоторых методах классификации и кластерного анализа. Однако всегда нужно помнить, что если модель строится с применением некоторых гипотез и предположений, то может оказаться, что они ошибочны.

Принятие решений в науке о данных базируется на результатах изучения и исследования данных в эмпирическом доказательстве. В науке о данных во всех ее областях и разделах внимание фокусируется именно на данных. Процессы принятия решений в ней базируются на тех гипотезах, моделях, новых знаниях, новых свидетельствах и фактах, которые получены в процессе работы с данными и на основе данных [1]. И это существенная методологическая черта науки о данных. Как отмечает автор [2], специалист в науке о данных исходит из данных, в отличие от других наук, работающих с данными, типа статистики, в которых исследователь больше заботится о методах. Последний гораздо больше концентрируется на решении вопросов типа «Какой алгоритм лучше и какой язык выбрать для программирования?», и проявляет гораздо меньше интереса к данным. В [2] также отмечается, что в науке о данных внимание к данным проявляется и в том, что касается ошибок в данных. Анализ влияния ошибок на конечный результат – это одна из базовых задач науки о данных. При работе с данными, содержащими ошибки и другие некорректности (выбросы, пропущенные значения и т.д.), специалисты других наук часто опираются на высказывание «мусор на входе – мусор на выходе»³. Понятно, что в окружающем мире нет ничего точного, абсолютно истинного или ложного. Позиция науки

о данных в этом вопросе состоит в том, что и с такими данными нужно работать.

4. Основные разделы науки о данных

Третий аспект любой науки (первые два – это свой объект исследований и своя специфическая методология, были описаны выше), которая претендует на самостоятельность – это направления исследований, которые отражают специфические научные аспекты науки. В настоящее время пока нельзя утверждать с уверенностью, что содержание науки о данных вполне сформировалось. Появляются новые типы данных (например, синтетические мультимедиа-данные), выборки данных с новыми свойствами, например, малые данные, которые ставят перед исследователями не менее сложные задачи, чем большие данные. Появляются новые идеи о том, как можно подходить к решению задач машинного обучения, и здесь примером являются генеративные сети. Наука о данных является пока еще молодой наукой, и по этой причине говорить о сложившейся тематике научных исследований, наверное, еще рано, тем более, что разные авторы имеют на этот счет различные мнения. Разные университеты могут иметь разные программы обучения на уровне аспирантуры, а на научных конференциях каждый год модифицируется список актуальных задач и новых приоритетов.

Авторское видение научной тематики науки о данных в настоящее время, которая должна отражать также и образовательный университетский курс для специалистов в области науки о данных на уровне магистратуры и аспирантуры, представлено ниже.

Раздел 1. Свойства данных и связи в данных.

1. *Источники данных, типы и базовые свойства данных.* Методы совместного использования данных, представленных в разных шкалах измерения, методы их совместного корректного агрегирования, грануляции и интеграции, в частности, корректные методы виртуализации гетерогенных данных, в том числе, зависящих от времени. Методы обогащения данных.

2. *Связи и отношения в данных, их свойства и допустимые преобразования.* Систематизация видов связей и отношений в данных, их теоретико-множественных, алгебраических и семантических свойств. Изучение структур данных разных типов с позиций абстрактной алгебры, а

³ Как тут не вспомнить высказывание Г. Гегеля: «Если факты противоречат моей теории — тем хуже для фактов». <https://ru.citaty.net/tsitaty/630301-georg-gegel-esli-fakty-protivorechat-moei-teorii-tem-khuzhe-dlia/>

именно систем с отношениями, порождающих систем, структур разных видов, их допустимых преобразований. Изучение гомоморфных преобразований структур гетерогенных данных в интересах корректного решения задач грануляции, интеграции и слияния (*англ. data fusion*) гетерогенных данных.

3. *Многомерные данные с различными видами неопределенности, их свойства и свойства отношений в данных с неопределенностью.* Ассоциации, корреляции и причинные связи в данных. Регрессионные зависимости. Причинные связи в данных с неопределенностью. Сети данных с неопределенностью. Вероятностные сети данных. Ассоциативные и причинные сети. Интеграция многомерных данных с разными типами неопределенности.

4. *Семантика данных, семантические отношения и семантические связи в данных.* Семантические отношения, квантификация отношений сходства. Семантическая близость, её роль в науке о данных и методы её измерения. Онтология как сеть семантических сущностей (понятий) и ее формальные свойства. Квантификация свойств онтологии и ее компонент. Соотношение семантики данных и семантики знаний.

Раздел 2. Сбор и повышение качества данных.

5. *Сбор и очистка данных, инжиниринг данных.* Интеллектуальные краулеры. Цифровые двойники как источники данных. Методы и алгоритмы очистки данных. Качество данных, пространство качества данных. Квантификация качества данных. Улучшение качества данных. Обнаружение ошибок и выбросов в данных. Восстановление пропущенных значений.

6. *Регуляризация и робастные вычисления с данными.* Устойчивость и вычислительная эффективность обработки больших данных на всех этапах их преобразования в процессах науки о данных. Доверие к данным. Избыточность данных и робастность вычислений.

7. *Агрегирование и грануляция данных.* Агрегирование данных. Грануляция данных. Теория грубых множеств и грануляция данных. Многоэтапная грануляция данных как семантическая технология принятия решений без традиционного машинного обучения.

Раздел 3. Пространства данных, их оптимизация и представление данных.

8. *Управление размерностью данных.* Размерность (многообразия) данных. Редукция размерности данных. Методы типа LASSO

снижения размерности больших данных. Методы случайных подпространств. Факторизация матриц как метод редукции пространств больших данных. Другие методы снижения размерности пространств данных. Проблемы робастности вычислений в задачах редукции размерности больших данных.

9. *Латентные пространства данных.* Понятие латентных переменных и латентных подпространств. Концептуальные основы построения латентных подпространств. Теоретические основы построения латентных пространств. Алгоритмы приближенного поиска латентных подпространств. Точные алгоритмы построения латентных подпространств и их эвристические упрощения. Методы поиска латентных подпространств в предметно-ориентированных приложениях.

10. *Визуализация многомерных данных* Новые концептуальные идеи и формы визуализации данных. Инструменты визуализации многомерных данных. Методы расширенной и виртуальной реальности для визуализации данных.

Раздел 4. От данных к знаниям и принятию решений.

11. *Новые методы извлечения связей и зависимостей в данных.* Новые методы предсказательной и предписывающей аналитики для сложных гетерогенных данных. Программные инструменты для поддержки процессов предсказательной аналитики.

12. *Новые задачи и методы кластерного анализа данных.* Кластерный анализ гетерогенных данных. Семантическая кластеризация данных. Средства интерактивной и автоматической разметки данных.

13. *Машинное обучение на основе новых принципов.* Инжиниринг признаков для решения задач машинного обучения. Новые принципы, методы, алгоритмы и инструменты машинного обучения для поддержки принятия управленческих решений, готовых к исполнению (*англ. actionable knowledge*). Новые методы обучения с подкреплением. Новые методы машинного обучения на основе гетерогенных данных. Методы многоцелевого обучения. Методы обнаружения знаний в мультимедиа-данных. Графовые и сетевые модели машинного обучения. Причинность и обучение причинных моделей. Онлайн-машинное обучение, потоковое, динамическое обучение и обучение в режиме реального времени. Многофакторное,

многоклассовое и многоцелевое машинное обучение. Обучение на основе частично размеченной выборки. Новые методы машинного обучения на основе малых данных. Методы и алгоритмы распределенного и р2р-обучения. Новые методы устойчивого машинного обучения в условиях атак на данные и на методы машинного обучения. Методы машинного обучения для обнаружения синтетического и манипулированного мультимедиа-контента. Онлайн-методы машинного обучения на основе данных социальных медиа.

14. *Поведенческие модели данных и групповое поведение.* Представление поведенческих знаний и модели группового поведения. Сценарные модели группового поведения. Новые задачи и методы машинного обучения классификации и предсказания группового поведения. Латентные функциональные пространства. Модели латентных переменных Гауссовских процессов. Машинное обучение на основе выборки нестационарных процессов. Нормальное и аномальное поведение.

15. *Экосистемы данных и знаний.* Единое информационное пространство данных и знаний. Компоненты инфраструктуры для поддержки единого информационного пространства. Семантические пространства данных и знаний.

Приведенный список задач и направлений исследований в области науки о данных – это пока набросок, который нуждается в уточнении, обсуждении и привлечении коллективного мнения. Со списком актуальных современных задач и направлений исследований в области науки о данных можно познакомиться на веб-сайте самой рейтинговой конференции в области науки о данных: DSAA-2022 [19]. Другие ресурсы науки о данных доступны на веб-сайте [20].

5. Проблемы и перспективы

Наука о данных пока остается молодой, а потому нерешенных проблем и вызовов в ней гораздо больше, чем достижений. Более того, несомненно, что многие важные проблемы, которые предстоит решать науке о данных, еще не осознаны. Это утверждение следует, с одной стороны, из актуальности данных как источника знаний для огромного числа научных и прикладных задач в области ИТ и ИИ, а с другой стороны – из роста сложности данных.

Очевидно, что проблемы и вызовы в науке о данных на современном этапе ее развития достойны специального исследования, и в рамки данной работы оно не укладывается. По этой причине далее только намечаются основные источники ключевых проблем науки о данных. Достаточно детальный анализ этих источников приведен в работе [1, гл. 4]. Материал этой работы частично используется далее в кратком анализе на эту тему.

Естественно, что основным источником алгоритмических проблем и вызовов являются сложности данных того или иного характера, того или иного происхождения. Если речь идет о больших данных, то нужно принимать во внимание, в первую очередь, дисбаланс между объемом данных и их размерностью, что ставит достаточно остро проблемы вычислительной сложности и робастности соответствующих алгоритмов. Эти проблемы особенно остро встают в случае, когда данные требуется обрабатывать в режиме онлайн и использовать результаты этой обработки в реальном времени или при работе с данными, обладающими высокой динамикой и частотностью, как это имеет место, например, в алгоритмическом трейдинге. Алгоритмические проблемы, причем связанные не только с вычислительной сложностью, возникают при работе с мультимедийными приложениями, при обработке гетерогенных данных из разных источников, при обработке данных с неопределенностью и сильно зашумленных данных. Неясно пока, как следует работать с данными, обладающими нечеткой структурой, в частности, нечеткой иерархией. Большие вычислительные проблемы характерны для работы с сильно разреженными данными и с неоднородными распределениями зависимых данных.

Другой источник не менее трудных проблем науки о данных – это сложные связи и отношения на множестве атрибутов данных, обычно скрытые от исследователя. Важность их поиска определяется хорошо известным фактом – данные в настоящее время рассматриваются как основной источник знаний, а знания определяются, в первую очередь, связями и отношениями на множестве их атрибутов. Сложные связи не всегда могут быть представлены известными их типами, например, ассоциациями, корреляциями, функциональными и/или причинными

связями. Предстоит большая работа по изучению их смесей, «явных и неявных связей, структурных и неструктурных отношений, семантических, иерархических и вертикальных отношений, а также эволюции отношений и рассуждений» [1, гл. 4 с. 94-95).

Гораздо больше новых проблем следует ожидать при работе с данными о поведении. Современная аналитика пока весьма слабо опирается на поведенческие данные ввиду их сложности. «...поведение отражает семантику и процессы, происходящие ... в реальном мире, которые часто игнорируются или существенно упрощаются после их преобразования существующими системами обработки данных в мир данных.» [1, гл. 4 с. 95). В результате многие важные свойства поведения в реальном мире не находят отражения в мире данных. По этой причине задачи предсказания решаются намного хуже, чем это потенциально возможно, если в алгоритмах предсказания явно использовать временные аспекты и сценарные модели бизнес- и других процессов. Например, поведение сложных объектов, состоящих из большого числа относительно автономных компонент, которое по существу является групповым поведением, невозможно предсказывать без учета фактора времени и сценарных моделей их поведения, в частности, в контексте особенностей предметной области приложения.

Социальные аспекты данных уже породили в науке о данных много новых проблем и вызовов, и этот аспект данных, возможно, уже в ближайшее время станет фокусом исследовательских усилий науки о данных. Уже сейчас появилось много новых сложных проблем этического характера, которые порождены социальными аспектами данных, и число таких проблем будет множиться.

Данные существуют не только в социальном контексте. Они существуют в контексте окружающей среды (такие сложности в науке о данных называют контекстными сложностями [1]). И здесь уже существует много проблем учета взаимодействия приложений с окружающей средой и контекстным поиском знаний для принятия решений. Однако появление наиболее сложных проблем следует ожидать в связи с тем, что решение многих задач науки о данных выполняется с вовлечением специалистов предметной области, с совместным использованием машинных и экспертных знаний.

Остановимся немного более подробно на проблеме качества данных, которая в настоящее время активно обсуждается как в научном, так и в бизнес-сообществах.

Качество данных – это их свойство, которое оценивает данные с прагматической точки зрения, а именно насколько они подходят для решения поставленной задачи, можно ли, используя конкретное множество данных, достигнуть цели, которая стоит перед исследователем. Это свойство данных должно поддаваться количественной оценке. Качество данных оценивается их корректностью, точностью, полнотой, достоверностью, согласованностью, целостностью, непротиворечивостью, надежностью и степенью доверия к ним [1]. Другой аспект измерения качества данных – это их оценки с позиций бизнес-целей, например, можно ли их использовать для улучшения качества бизнеса, для стратегического планирования, или, например, принятия тактических решений по управлению бизнесом и поддержки принятия решений в интересах бизнеса.

Данные – это очень ценный, если не самый ценный актив любой организации и компании, причастной к решению проблем ИТ и ИИ. Понимание этого факта определяет отношение исследователей и разработчиков к проблемам качества данных и технологиям его обеспечения. По этой причине анализ качества данных, методы и средства его улучшения – это одна из актуальных задач науки о данных, поскольку данные – это в настоящее время основной источник знаний для приложений ИИ, а знания во многом определяют их возможности.

Исследователи в области науки о данных сходятся во мнении, что контроль качества данных начинается при их сборе и продолжается в течение всего их жизненного цикла. Опыт большинства компаний подтверждает, что 80% времени работы с данными уходит на их очистку и повышение качества, и только 20% – на все остальное, что относится к жизненному циклу данных. Этот опыт важен, и он говорит о серьезности проблемы и о ее месте в общем множестве проблем и вызовов науки о данных. Практики говорят, что при работе с источниками данных, которые оказываются плохими, нужно заниматься, прежде всего, не постоянной очисткой данных, а поиском и ликвидацией причин, которые делают данные «плохими». Этот вариант того, что в общем случае часто

называют обратным инжинирингом, требует научного обоснования. В процессе поиска и анализа ошибок в данных очень важно предсказывать проблемы, которые эти ошибки будут вызывать при последующем практическом использовании данных. И это тоже серьезная научная проблема.

Особую важность имеет проблема профилирования данных, которая имеет целью описать, проанализировать и определить тематику, существенные признаки, характеристики, отношения, структуры и другие стороны и свойства выборки данных. В этой связи необходимо тщательно документировать данные и продолжать документировать данные в течение всего жизненного цикла их создания, представления, хранения и использования. Многие компании имеют свои отработанные вопросники по качеству данных и правила ее проверки, но эти вопросники и правила анализа качества данных обычно относят к коммерческой тайне и не разглашают. В эпоху открытых данных, открытой науки, открытого программного кода и открытого образования это смотрится как архаизм. Естественно, что при создании таких вопросников нужно иметь научное обоснование их полноты и непротиворечивости.

Большие проблемы, связанные с качеством данных, возникают в случае сетей данных (*англ.* Data Mesh), т.е. данных с распределенной архитектурой хранения и с централизованным управлением, со стандартами, поддерживающими технологию интеграции данных, и с инфраструктурами, поддерживающими работу с такими данными.

Проблема качества данных – это не проблема ИТ, а проблема науки о данных, в которой большую роль играют организационные аспекты.

В целом, проблема анализа и повышения качества данных пока еще находится на стадии становления, и в настоящий момент остается скорее ремеслом, постигаемым на практике, чем научно обоснованной технологией.

Приведенный краткий анализ источников проблем в науке о данных свидетельствует о широком фронте исследований, которые предстоит выполнить в процессе развития науки о данных от ее современного состояния до такого, когда науку о данных можно будет считать достаточно зрелой.

Заключение

Данная работа дает краткое введение в существо науки о данных, в ее методологические основы и научные направления, а также объясняет причины быстрого роста ее популярности в научном и индустриальном сообществах, специализирующихся в области исследований и разработок интеллектуальных систем. Эта наука за полтора десятилетия своего существования достигла определенной зрелости, обрела самостоятельность и заняла важное место среди других наук, которые занимаются проблемами обработки данных.

В отличие от других наук, которые так или иначе используют данные для решения различных задач, наука о данных занимается всесторонним изучением данных, методологическими и алгоритмическими аспектами современных технологий сбора, хранения, подготовки, представления и использования данных в интересах обнаружения знаний и их преобразованием в решения и в действия по исполнению принятых решений. Наука о данных сделала доступными новые источники знаний, необходимых для вычислительно эффективной и устойчивой работы интеллектуальных систем в самых различных классах приложений. По существу, именно наука о данных придала мощный импульс современному развитию ИИ, повышению интереса к использованию интеллектуальных приложений и доверия к ним со стороны индустриального сообщества.

Методология решения задач в науке о данных сильно отличается от традиционных методологий использования данных в научных исследованиях и прикладных разработках. Эта методология рассматривает данные как основной источник знаний и решений. При извлечении знаний эта методология не опирается на какие-либо гипотезы, а ставит целью получение гипотез о свойствах данных, ожидаемых результатах и итоговой ценности этих результатов на основе исследования самих данных. Методологические принципы науки о данных не предполагают делать априорный выбор модели знаний и модели принятия решений, когда научная задача сводится к поиску значений конкретных параметров, выбором значений которых результирующая модель знаний и принятия решений подгоняется под имеющиеся

данные с помощью тех или иных методов оптимизации. Методология науки о данных опирается на выводы на основе обнаруженных фактов и считает эмпирическое доказательство столь же правомерным и убедительным, как и теоретическое доказательство того или иного вида.

Эти черты методологии науки о данных являются чрезвычайно важными, поскольку именно они существенно расширяют границы применимости численных исследований на основе данных, делают их применимыми в таких науках, как общественные и политические науки, как область истории, культуры и искусства, в которых ранее численные методы никогда не применялись.

Наука о данных показала, что данные пока еще изучены слабо, что они являются источником знаний и ценности, о которых ранее не было известно. Можно с уверенностью утверждать, что дальнейшие исследования и разработки в области науки о данных приведут к инновационным результатам в самых разных областях науки и общественной жизни, откроют новые возможности и перспективы.

Литература

- Longbing Cao. Data Science Thinking. The Next Scientific, Technological and Economic Revolution. International Publishing AG, part of Springer Nature. 2018. 367 p.
- Steven S. Skiena. The Data Science Design Manual. Texts in Computer Science, Springer, 2017, 445 p. (русский перевод Стивен С. Скиена. Наука о данных. Учебный курс. Вильямс. 2020. 544 с.
- Грас Джоэл. Data Science. Наука о данных с нуля: Пер. с англ. СПб. БХВ-Петербург, 2017. 336 с.
- Силен Дэви, Мейсман Арно, Али Мохамед. Основы Data Science и Big Data. Python и наука о данных. ООО Питер Пресс. 2017, 336 с.
- Келлехер Джон, Тирни Брендан. Наука о данных. Альпина Диджитал, 2018, 143 с.
- Laura Igual and Santi Seguí. Introduction to Data Science. Undergraduate Topics in Computer Science, Springer International Publishing Switzerland. 2017. 218 p.
- The Field Guide to Data Science. Booz/Allen/Hamilton Inc., 2nd Edition. 2015. URL <https://wolfpaulus.com/wp-content/uploads/2017/05/field-guide-to-data-science.pdf>
- Городецкий В.И., Юсупов Р.М. Искусственный интеллект: метафора, наука и информационная технология. Мехатроника, автоматизация, управление. 2020, 21(5). 282 – 293. URL <https://doi.org/10.17587/mau.21.282-293>.
- Russell Ackoff. From data to wisdom, Journal of Applied Systems Analysis, 1989, 16, 3 – 9. URL <http://www-public.imtbs-tsp.eu/~gibson/Teaching/Teaching-ReadingMaterial/Ackoff89.pdf>
- Wikipedia. Dikw pyramid. URL https://en.wikipedia.org/wiki/DIKW_Pyramid
- Chaim Zins. Conceptual Approaches for Defining Data, Information, and Knowledge. Journal of the American Society for Information Science and Technology. 2007, 58(4), 479 – 493. URL http://www.success.co.il/is/zins_definitions_dik.pdf
- Jennifer Rowley. The wisdom hierarchy: representations of the DIKW hierarchy. Journal of Information Science. 2007, 33(2), 163–180. URL <http://web.dfc.unibo.it/buzzetti/IUcorso2007-08/mdidattici/rowleydikw.pdf>
- Кузнецов О.П. Когнитивная семантика и искусственный интеллект. Искусственный интеллект и принятие решений, № 4. 2012. С. 32 – 42.
- Jianqing Fan, Fang Han, and Liu Han. Challenges of Big Data Analysis.// Princeton University, Johns Hopkins University, August 7, 2013. URL <http://arxiv.org/pdf/1308.1479.pdf>
- Городецкий В.И. Проблемы обработки больших данных //XIV национальная конференция по искусственному интеллекту с международным участием КИИ-2014 (24-27 октября 2014 г. Казань, Россия): Труды конференции в 3-т., Казань: Изд-во РИЦ «Школа». 2014. Т.2. С. 231 – 242. URL <http://raai.org/resurs/papers/cai/2014/cai2014-vol2.pdf>
- Jianqing Fan, Runze Li, Cun-Hui Zhang, Hui Zou. Statistical Foundations of Data Science. Chapman and Hall/CRC. 2020. 774 p.
- Daniel Peña, Ruey S. Tsay. Statistical Learning for Big Dependent Data. Wiley Series in Probability and Statistics. 629 p.
- James D. Miller. Statistics for Data Science. Leverage the power of statistics for Data Analysis, Classification, Regression, Machine Learning, and Neural Networks. 2017. Packt Publishing Ltd. 286 p.
- IEEE International Conference on Data Science and Advanced Analytics. URL <http://dsaa2022.dsaa.co/http://dsaa2022.dsaa.co/>
- Ресурсы науки о данных. URL <https://datasciences.info/>

Городецкий Владимир Иванович. Доктор технических наук, профессор. Ведущий научный сотрудник АО «ЭВРИКА». Области исследований: прикладная статистика, планирование и составление расписаний, искусственный интеллект и принятие решений, многоагентные системы, групповое управление, самоорганизующиеся сети агентов, инструментальные системы, машинное обучение, эффективные и робастные алгоритмы обработки больших данных, семантические вычисления, распределенное и P2P обучение, компьютерная безопасность, системы с неопределённостью, стеганография, модели коллективного поведения роботов и др. E-mail: vladim.gorodetsky@gmail.com

Data Science: Methodology, Main Directions, Problems and Perspectives

V. I. Gorodetsky

Joint-stock company "EURECA", St. Petersburg

Abstract. Data science as a special science has appeared in the early 2000s. Its appearance was a kind of response of the scientific community to the challenges of big data. It has now reached a level of maturity that makes it a science, playing an important and indispensable role in the cluster of information sciences dealing with the problems of data collection, storage, presentation and use. The paper outlines the boundaries of data science in relation to artificial intelligence. It describes the multiple two-way relationships of data science and other related sciences dealing with data, and in brief introduces to the data science methodology and its main scientific fields. Finally, the paper indicates the main sources of challenges and difficult problems those are expected to be solved by the data science.

Keywords: data science, artificial intelligence, DIKW pyramid, data science thinking, data science challenges.

DOI 10.14357/20718594220301

References

1. Longbing Cao. Data Science Thinking. The Next Scientific, Technological and Economic Revolution. International Publishing AG, part of Springer Nature. 2018. 367 p.
2. Steven S. Skiena. The Data Science Design Manual. Texts in Computer Science, Springer, 2017, 445 p. (русский перевод Стивена С. Скиена. Наука о данных. Учебный курс. Издательство Вильямс. 2020. 544 p.
3. Джозел Грас. Data Science. Наука о данных с нуля: Пер. с англ. СПб. БХВ-Петербург, 2017. 336 p.
4. Дэви Силен, Арно Мейсман, Мохамед Али. Основы Data Science и Big Data. Python и наука о данных. ООО Питер Пресс. 2017, 336 p.
5. Джон Келлехер, Брендан Тирни. Наука о данных. Альпина Диджитал, 2018, 143 стр.
6. Laura Igual and Santi Seguí. Introduction to Data Science. Undergraduate Topics in Computer Science, Springer International Publishing Switzerland, 2017, 218 pp.
7. The Field Guide to Data Science. Booz|Allen|Hamilton Inc., 2nd Edition, 2015, URL <https://wolfpaulus.com/wp-content/uploads/2017/05/field-guide-to-data-science.pdf>
8. В.И. Городецкий, Р.М. Юсупов. Искусственный интеллект: метафора, наука и информационная технология. Мехатроника, автоматизация, управление. 2020, 21(5), 282 – 293. URL <https://doi.org/10.17587/mau.21.282-293>,
9. Russell Ackoff, From data to wisdom, Journal of Applied Systems Analysis, 1989, 16, 3 – 9. URL <http://www-public.imtbs-tsp.eu/~gibson/Teaching/Teaching-ReadingMaterial/Ackoff89.pdf>
10. Wikipedia. Dikw pyramid. URL https://en.wikipedia.org/wiki/DIKW_Pyramid
11. Chaim Zins. Conceptual Approaches for Defining Data, Information, and Knowledge. Journal of the American Society for Information Science and Technology. 2007, 58(4), 479–493. URL http://www.success.co.il/is/zins_definitions_dik.pdf
12. Jennifer Rowley. The wisdom hierarchy: representations of the DIKW hierarchy. Journal of Information Science. 2007, 33(2), 163–180. URL <http://web.dfc.unibo.it/buzzetti/IUcorso2007-08/mdidattici/rowleydikw.pdf>
13. О.П. Кузнецов. Когнитивная семантика и искусственный интеллект. Искусственный интеллект и принятие решений, № 4, 2012. 32 – 42.
14. Jianqing Fan, Fang Han, and Liu Han. Challenges of Big Data Analysis.// Princeton University, Johns Hopkins University, August 7, 2013. URL <http://arxiv.org/pdf/1308.1479.pdf>
15. В.И. Городецкий. Проблемы обработки больших данных //XIV национальная конференция по искусственному интеллекту с международным участием КИИ-2014 (24-27 октября 2014 г., Казань, Россия): Труды конференции в 3-т., Казань: Изд-во РИЦ «Школа», 2014, Т.2, 231 – 242. URL <http://raai.org/resurs/papers/cai/2014/cai2014-vol2.pdf>
16. Jianqing Fan, Runze Li, Cun-Hui Zhang, Hui Zou. Statistical Foundations of Data Science. Chapman and Hall/CRC. 2020, 774 pp.
17. Daniel Peña, Ruey S. Tsay. Statistical Learning for Big Dependent Data. Wiley Series in Probability and Statistics, 629 pp.
18. James D. Miller. Statistics for Data Science. Leverage the power of statistics for Data Analysis, Classification, Regression, Machine Learning, and Neural Networks. 2017. Packt Publishing Ltd, 286 pp.
19. IEEE International Conference on Data Science and Advanced Analytics. URL <http://dsaa2022.dsaa.co/http://dsaa2022.dsaa.co/>
20. Ресурсы науки о данных. URL <https://datasciences.info/>

Gorodetsky Vladimir I. Doctor of technical sciences, professor. Leading researcher, Joint-stock company "EURECA". Research areas: planning, pattern recognition and artificial intelligence, multi-agent systems, knowledge discovery from data and data fusion, p2p agent-based systems and technologies, self-organizing systems, computer network security, distributed data mining and knowledge discovery, dealing with uncertainty, steganography, collective robotics, etc. E-mail: vladim.gorodetsky@gmail.com