

Лонбин Цао¹**Образ мышления в науке о данных:**

Наступающая научно-техническая и экономическая революция.

СПб. : Издательство Европейского университета в Санкт-Петербурге. 2022. 552 с.

Springer International Publishing AG, 2018. Перевод с английского А.В. Климонтович.

Научный редактор В.И. Городецкий

Современный быстроразвивающийся мир данных, интенсивное использование данных и новые возможности научного поиска и бизнеса на их основе породили науку о данных — новую парадигму исследований и разработок, которая использует методы и возможности экспериментальной, теоретической и вычислительной науки. Как любая новаторская область знания, наука о данных порождает много споров и вопросов: что именно она из себя представляет, как проникает в различные профессии, образование, бизнес, экономику и другие сферы нашей жизни, как конкурирует в этих областях с другими науками и как их изменяет. В этой книге представлена по возможности полная картина науки о данных как новой научной и технологической парадигмы, а также учебной дисциплины. Автор подробно рассказывает об эволюции образа мышления, основанного на данных, о развитии входящих в науку о данных дисциплин, о том, как и почему эта наука становится движущей силой новой цифровой экономики. В книге описывается, как мышление на основе данных трансформирует наши представления об управлении и стратегии развития предприятия, как оно формирует прорывные научные исследования и технологические инновации. Книга может быть полезна для менеджеров данных, бизнес-аналитиков, лиц, принимающих стратегические решения, руководителей исследований и преподавателей, ответственных за продвижение актуальной научной, инновационной и промышленной повестки и за разработку учебных курсов следующего поколения, а также для всех, кто хочет понять, как наука о данных меняет наш мир.

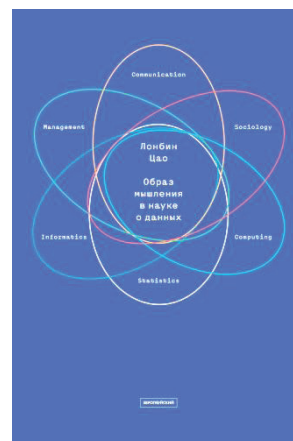
В книгу включен дополнительный материал: толковый англо-русский словарь терминов науки о данных, подготовленный профессором В.И. Городецким².

Фрагмент толкового англо-русского словаря терминов науки о данных³

Analytics. Аналитика, аналитика данных: Раздел науки о данных, включающий множество теорий, технологий, инструментальных средств и процессов, которые дают возможность глубоко понять и эффективно выявить практически полезные гипотезы, зависимости и знания, скрытые в данных. Аналитическая обработка данных включает дескриптивную аналитику, предсказательную и предписывающую аналитику.

Artificial Intelligence. Искусственный интеллект (ИИ): Наука о получении, представлении и использовании знаний. Её развитие опирается в основном на информатику, вычислительную технику, машинное обучение, теорию взаимодействия и теорию систем, использует вычислительную логику, планирование, обучение с подкреплением, статистическое/вероятностное представление, рассуждения на основе прецедентов, эволюционные вычисления и глубокий поиск. Эти науки покрывают большую (но не всю) часть существующих частных направлений исследований по ИИ. Их краткий список включает эмуляцию памяти, извлечение информации, её кодирование, хранение, индексацию и поиск, машинное обучение, обнаружение связей и закономерностей в данных, распознавание речи, языка, голоса, изображений и мультимедийных материалов, нейросетевые вычисления, генеративные нейросети, нечёткие и эволюционные вычисления, цифровые двойники объектов и др.

Domain data intelligence. Интеллект предметной области: Понятие, которое характеризует знания, интегрированные в предметную область (ПрО), к которой относятся используемые данные или системы данных. Характеризует значимые свойства ПрО и их понимание, появляется в контексте свойств ПрО. Контекст может быть представлен релевантными атрибутами описания ПрО, знаний и метазнаний о ней, а также другими ресурсами, специфичными для ПрО. Включение качественных и количественных знаний о ПрО может способствовать более глубокому пониманию её сложностей и их ключевой роли в выявлении неизвестных знаний и решений из данных о ПрО в форме рекомендуемых действий и механизмов принятия решений, например в области биоинформатики или социологических вычислений.



¹ Лонбин Цао (р. 1969) — бизнес-аналитик, профессор информационных технологий в Сиднейском технологическом университете (Австралия), главный редактор журнала *International Journal of Data Science and Analytics*.

² Продолжается дискуссия о содержании эволюционирующих терминов в области, связанной с данными, обработкой данных, ..., искусственным интеллектом. Редакция журнала поддержала профессора В.И. Городецкого, внёсшего свой вклад в терминологический спор и предложившего своё толкование важнейших терминов, и надеется на конструктивное их обсуждение читателями журнала.

³ Толковый словарь составлен профессором В.И. Городецким в основном по материалам перевода книги Longbing Cao. *Data Science Thinking. The Next Scientific, Technological and Economic Revolution*. International Publishing AG, part of Springer Nature 2018. Смысл некоторых терминов обсуждался, в том числе, и с автором этой книги Longbing Cao.

Environment data intelligence. *Интеллект окружающей среды:* Интеллект, скрытый в данных о внешнем окружении проблем и задач науки о данных. Факторы окружающей среды могут быть описаны данными в терминах ПрО, организационных, социальных, человеческих, сетевых и/или других связанных с ними понятий. Факторы окружающей среды в задаче о данных или в системе данных зависят от ПрО и конкретной задачи. Кроме предметно-зависимых контекстуальных факторов имеются общие моменты, связанные с окружающей средой, например взаимодействия и отношения внутри контекста, взаимодействия и отношения между контекстом и системой данных, влияние контекстуальных факторов и отношений на систему данных, динамика и эволюция среды и её воздействие на задачу. Эти факторы важны для понимания данных о среде и манипулирования ими.

Data science, datalogy. *Наука о данных, даталогия:* Термины являются синонимами, хотя последний вариант используется редко даже в зарубежном сообществе. Можно ожидать, что со временем для науки о данных будет принят более краткий термин даталогия. Это грамматический и отчасти смысловой аналог названий ряда других наук, например биологии.

Human intelligence (in data science). *Человеческий интеллект (в науке о данных):* Интеллект отдельной личности в контексте науки о данных. Это интеллект, который привносится человеком в процесс решения проблем, связанных с данными, или в исследование систем данных. При решении сложных задач науки о данных человеческий интеллект играет ключевую роль в понимании обрабатываемых данных и их контекста, в проектировании и реализации решений, а также в выявлении ценности данных. Человеческий интеллект в науке о данных может иметь форму явного привлечения эмпирических знаний людей, их мнений, намерений, ожиданий, форму явного привлечения человека к работе с данными в реальном времени, использования оценок отдельных экспертов или их групп при рассмотрении сложных проблем науки о данных. Он может иметь форму неявного или косвенного привлечения образного мышления человека, его рассуждений, основанных на воображении, привлечения эмоционального интеллекта человека, его вдохновения, способности к мозговому штурму, умения рассуждать и познавать с помощью конвергентного мышления посредством взаимодействия с другими членами команды при решении различных проблем науки о данных. В зависимости от уровня сложности и предъявляемых требований человеческий интеллект может выполнять разные роли в науке о данных.

Human-like machine intelligence. *Человекоподобный ИИ:* Интеллект, ориентированный на разработку новых механизмов в ИИ современного уровня. Ключевые моменты человеческого интеллекта, которые в настоящее время плохо моделируются в ИИ, но которые следовало бы включить в следующее поколение систем ИИ, а именно: интуиция, энтузиазм, любознательность, мышление на основе воображения, вдохновение, креативное мышление, иерархическая сложность, иерархический интеллект и их взаимосвязи. Наиболее трудная задача современных исследований в области человекоподобного ИИ — понять, как можно выполнить моделирование работы человеческого мозга и как можно реализовать человеко-подобный мыслительный процесс.

Network data intelligence. *Интеллект данных сети (сетевой интеллект):* Интеллект, интегрированный в сеть, которая используется для решения задач, связанных с данными или с системой данных. Сетевой интеллект возникает из веб-интеллекта или из широкомасштабных сетевых процессов и сетевых ресурсов, а также из активности пользователей, например, в социальных сетях, при работе с мобильными сервисами, сетевыми системами и при работе с другими живыми существами. Сетевой интеллект реализуется в таких аспектах, как распределение информации и ресурсов по сети, представление и учёт связей между распределёнными объектами и системами, отношения и взаимодействия между сетевыми узлами, скрытые сообщества и группы, формирующиеся в сети, информация и другие ресурсы, о которых становится известно из сетей. Сетевой интеллект содержится в веб-ресурсах, облачной инфраструктуре и облачных вычислительных ресурсах, в информации, извлекаемой средствами поиска, в способностях к структуризации данных распределённых хранилищ и внешней среды.

Knowledge. *Знание:* Представляет собой форму обработанной информации в терминах смеси информации, процедурно описанных действий или пропозициональных правил. Знание может быть субъективным или объективным, известным или неизвестным, применимым на практике или нет и разумным или нет. См. место понятия «знание» в пирамиде DIKIW.

Wisdom. *Мудрость:* Представляет собой высокоуровневые правила, которые являются продуманным результатом обработки информации, знаний или просто результатом вдохновения, интуиции, приобретённых из опыта или в процессе интеллектуальной деятельности. Мудрость указывает на высокие способности её носителя, на его метазнания, разум, умение применять знания на практике, на его способность к суждениям или принятию правильных решений в нужном месте, в нужное время и для нужных целей. Мудрость может быть нематериальной, уникальной, личной, интуитивной или может быть результатом умственного вдохновения. По сравнению со знанием, мудрость находится вне времени. Она является всесторонней, общей и эмоциональной, передаётся из поколения в поколение и между культурами.

Data DNK. *ДНК данных:* Даталогическая «молекула» данных, состоящая из фундаментальных и типовых элементов: поведение (англ. *behavior*), сущность (англ. *entity*), отношение (англ. *relationship*) и свойство (англ. *property*). В науке о данных ДНК данных приписывают роль, аналогичную той, которую биологическая ДНК играет в живых организмах. Четыре элемента в ДНК данных, а именно поведение, сущность, отношение и свойство (BERP) представляют различные фундаментальные аспекты данных. Сущность может быть объектом, примером (экземпляром), человеком, организацией, системой или частью подсистемы или окружающей среды. Свойство касается атрибутов, которые описывают сущность. Поведение описывает действия или динамику сущности или набора сущностей. Отношение описывает взаимодействие сущностей и свойств, включая взаимодействие значений свойств.

DIKIW. (от англ. *Data, Information, Knowledge, Intelligence, Wisdom*): DIKIW: Сокращённое название когнитивной пирамиды обобщения данных по уровням Данные — Информация — Знания — Интеллект — Мудрость.

Information (in data science). *Информация:* Представляет собой описание данных (объектов) в организованном виде с определённой целью или в виде, имеющем определённый смысл. Информация может быть структурированной (организованной) или функциональной (целенаправленной), субъективной или объективной (основанной на фактах) и т. д. Обычно, чтобы подчеркнуть отличие информации от данных, говорят, что информация — это семантическое восприятие данных человеком с позиций её содержания, представления и использования. См. также место понятия «информация» в пирамиде DIKIW.

Data. *Данные:* Дискретные или непрерывные факты, сигналы (например, датчиков, которые могут быть субъективными или объективными) или символы (знаки), которые относятся к объекту (к физической или виртуальной сущности или к событию). Данные лежат на нижнем уровне когнитивной системы DIKIW и могут быть субъективными или объективными, иметь или не иметь смысл и ценность.

Brain informatics. *Информатика мозга:* Научное направление в области исследований мозга человека и живых организмов, которое использует методы и модели информатики и искусственного интеллекта для изучения и исследования процессов и механизмов, используемых мозгом человека и живых организмов для решения своих задач, в частности задач принятия решений.

Computational intelligence. *Вычислительный интеллект:* Направление в ИИ, которое использует аналоги некоторых механизмов принятия решений человеком и живыми существами, в частности для решения отдельных задач науки о данных. Включает методы и технологии нейронных сетей, нечёткой логики и эволюционных вычислений. Эти технологии способны решать ряд специфических задач науки о данных лучше, чем другие того же назначения.

Data coupling. *Связи в данных:* Любые отношения между данными, например совместная встречаемость, пространственное соседство, функциональная зависимость, взаимодействие, ассоциация, корреляция, причинность, сходство, упорядоченность во времени, синонимия и т. п.

Data engineering. *Инженерия данных:* Относится к процессам извлечения, поиска, сбора, подготовки, предварительной обработки и управления данными. Выполняется инженерами по данным в интересах специалистов по анализу данных.

Data intelligence. *Интеллект данных:* Это знания, встроенные в данные (скрытые в данных). Могут быть выявлены с помощью методов машинного обучения. Критерием успешности выявления интеллекта данных является то, в каком объёме и до какой степени специалист по данным сможет понять, осмыслить и уловить свойства данных, их сложности и потенциал.

Raw data. *Сырые (необработанные) данные:* Данные, которые ещё не обработаны или не готовы для использования. Необработанные данные называют также первичными данными.

Data things. *Вещи данных:* Понятие, которое используется в экономике данных. Оно отличается от физических объектов (вещей) традиционной экономики. Вещи данных — это самоорганизующиеся объекты, которые сами обеспечивают себя необходимыми данными (англ. *self-datafied*), виртуализируются и устанавливают связи между собой, формируя новые вещи данных более высокого уровня агрегирования и фабрики вещей данных.

Data products. *Информационные продукты:* Информация, знания, программные системы, программные инструменты и т. д., которые являются результатами использования методологий, методов, алгоритмов и инструментов науки о данных. Являются результатами обработки данных или получаются с использованием данных, могут быть представлены множеством найденных свойств, предсказаний значений переменных, состояний системы или событий, происходящих в ней. Могут иметь форму сервисов, рекомендаций, решений для исполнения, извлечённых знаний, моделей, алгоритмов, новых парадигм, систем или приложений.