

А.И. Пестунов, А.С. Гинтофт, О.В. Криветченко

Big data как феномен: причины и следствия появления больших данных

УДК 330.46

Аннотация. В статье показано, что многие технологии, называемые теперь технологиями больших данных, существовали задолго до внедрения в массовый обиход данного термина. Принципиальное отличие нынешней ситуации состоит лишь в масштабе использования и некоем синергетическом эффекте, возникшем в результате развития целого ряда изначально слабо связанных технологий. Помимо технико-экономических, поднимаются гуманитарно-этические и социальные проблемы. Названы профессии, которые становятся актуальными в рамках технологий больших данных и востребованные в этой связи цифровые компетенции. По-новому ставится проблема защиты персональных данных.

Ключевые слова: большие данные; управление на основе данных; цифровая экономика; цифровые компетенции; искусственный интеллект; машинное обучение; нейронная сеть; демон Лапласа; детерминизм; персональные данные; этика; мораль; социальные противоречия цифровизации

Введение

Одной из сквозных технологий в проекте «Национальная технологическая инициатива» являются «большие данные»¹. Этот термин наряду с англоязычным вариантом Big Data приобрел широкое распространение в различных сферах, однако однозначного понимания, что это такое, до сих пор не выработано. Каким образом просто «данные» стали «большими» и чем они отличаются? Может быть, введение этого понятия – лишь маркетинговый ход, чтобы продать массовому клиенту привычные товары и услуги дороже? Например, еще в 2012 г. существовало несколько определений больших данных, ни одно из которых в полной мере не может удовлетворить исследователей, поскольку понятия, используемые в этих определениях, допускают неоднозначные интерпретации [Floridi, 2012].

¹ Национальная технологическая инициатива [Эл. ресурс]. URL: <https://nti2035.ru/nti> (дата обращения: 04.05.2023).

Скорее всего, подобные проблемы с интерпретацией сохранятся и в ближайшие годы. Вряд ли какое-то определение «больших данных» будет единодушно принято научно-технологическим сообществом. Поэтому на данном этапе имеет смысл рассматривать их не как понятие, а как феномен, заставляющий взглянуть на некоторые существующие объекты иначе и выявить их новые свойства.

В литературе нередко свойства больших данных описываются английскими прилагательными, начинающимися на литеру «V», из-за чего их часто называют V-характеристиками. Базовые свойства принято обозначать моделью «3V»: volume (объем), velocity (скорость) и variety (разнообразие). Если добавить к ним veracity (достоверность), получится модель «4V». Аналогичным образом можно добавлять, например, value (ценность), variability (изменчивость) и visualization (визуализация), формируя модели «5V», «7V» и т.д.

Особенность этих свойств состоит в том, что, во-первых, некоторые из них неизмеримы, и, во-вторых, не существует точных границ, определяющих тот момент, когда данные приобретают то или иное свойство. Более того, иногда складывается впечатление, что перечень V-характеристик разрастается неоправданно. Фактически инициаторы берут слово, начинающееся на «V», и пытаются обосновать причину его включения в список.

Технологии, называемые «большими данными», используются как в естественно-научных областях (например, в химическом анализе [Мильман, Журкович, 2020], в науках о Земле [Гвишиани и др., 2022]), так и в экономике и управлении [Булгаков, 2022; Засухина и др., 2022; Рыльникова и др., 2022; Половинченко, Елисеев, 2021]. При этом не всегда явно видно, какие принципиально новые подходы и методики стали использоваться и почему следует выделять технологии больших данных в отдельный блок.

Настоящая статья представляет собой попытку разобраться, как данные стали большими, и кто может использовать технологии больших данных для повышения эффективности своей деятельности. Big data при этом рассматриваются не как конкретная технология, а как феномен, о котором вдруг многие стали говорить как о некоем событии, разделившем мир на «до» и «после».

Желание стать «демоном Лапласа» как драйвер развития технологий больших данных

Принятие решений, как правило, заключается в попытках спрогнозировать развитие событий в условиях неопределенности и выбрать наиболее благоприятный вероятный вариант. Их качество во многом зависит от качества исходной информации. Мы стремимся, чтобы она была достоверной, полной и хорошо представленной для работы с ней.

Если рассмотреть идеальную ситуацию, когда наша информация является исчерпывающей, не вызывающей сомнений и максимально доступной для восприятия, акт принятия решения фактически сводится к выбору того варианта развития событий, который лучше соответствует заданной цели.

В пределе, когда некий субъект владеет всей полнотой информации, получается ситуация философского детерминизма, базовый тезис которого гласит: «Если бы какое-нибудь разумное существо смогло узнать положение и скорость всех частиц в мире, оно могло бы совершенно точно предсказать все события Вселенной». Такое гипотетическое существо с феноменальной осведомленностью о прошлом и настоящем, которое имеет абсолютные средства использования этой информации для предсказания и управления будущим, названо «демоном Лапласа» по фамилии ученого, предложившего этот мысленный эксперимент.

Таким образом, чем больше информации собрано и структурировано, тем сильнее ее владелец становится похож на демона Лапласа. На практике это означает, что чем дальше развиваются технологии сбора, обработки и хранения информации, тем лучше люди могут прогнозировать и, следовательно, принимать более эффективные решения. Но именно эти технологии в последние десятилетия совершили огромный скачок, создав великий соблазн стать демоном Лапласа. Благодаря этому скачку данные превратились из обычных в «большие», а технологии, позволяющие воплотить этот соблазн в жизнь, являются технологиями больших данных.

Автоматизация сбора данных как основная причина появления Big Data

Причиной поступления данных о наблюдаемом объекте являются всевозможные действия и изменения состояния (например, перемещение в пространстве или изменение температуры, давления и других параметров). Эта информация может фиксироваться

для хранения и передачи путем записи на разных носителях (от камня и бумаги до диска компьютера).

Важно отметить, что все эти действия совершались на протяжении всего существования человечества и мира в целом. Люди зарисовывали сцены охоты на стенах пещеры, составляли морские карты и атласы движения звезд, писали хроники и энциклопедии и т.д. Но почему большие данные возникли только в последние десятилетия?

Основная причина – это появление технологической возможности автоматизировать весь процесс сбора и обработки информации. Сначала люди научились регистрировать и хранить данные практически обо всех заметных событиях в машиночитаемом (т.е. пригодном для дальнейшей автоматической обработки) формате. Затем появились относительно недорогие датчики, видеокамеры и измерительные приборы, способные снимать показания без участия оператора и передавать их к месту хранения. В настоящее время ведется работа по цифровой трансформации всех сфер деятельности, в рамках которой меняются многие бизнес-процессы, внедряются автоматизированные системы управления, позволяющие формировать и передавать командные сигналы на основе полученной информации. Наконец, произошел резкий скачок объема систем хранения данных – появились дата-центры.

Важно, что весь процесс работы с информацией может быть практически полностью автоматизирован. Там, где раньше требовался оператор или стенографист, обладавший крайне ограниченной производительностью, теперь все записывается автоматически, и ровно с той скоростью, с которой протекает процесс.

Известный тезис о том, что за последние пару десятилетий получено данных больше, чем за всю историю человечества, следует трактовать в том смысле, что речь в нем идет именно о зарегистрированных (записанных) данных. Изменения происходили всегда, и многие из них фиксировались на материальных носителях. Разница в том, что теперь все происходящие изменения и события оставляют цифровой след.

Возможность регистрации и обработки огромных массивов информации в режиме реального времени ассоциирует феномен больших данных с тремя наиболее важными V-характеристиками – это *объем* (volume), *разнообразие* (variety) и *скорость* (velocity).

Примечательно, что ни одно из этих качеств не может быть определено точно. Сложно указать тот объем, при достижении

которого просто данные становятся «большими». Аналогично невозможно определить те параметры структуры данных, когда они начинают обладать свойством разнообразия. Скорость тоже плохо поддается точному описанию. Одним словом, все эти характеристики не столько описывают феномен, сколько акцентируют внимание на качественных изменениях данных.

Так, для обычного пользователя данные становятся «большими», если их объем выходит за рамки ограничений базовых компьютерных программ, таких как Excel (не более 1 млн строк) или Google-таблицы (не более 10 млн ячеек). В то же время для специалиста, использующего другие инструменты и программы, ограничения будут иными. Для кого-то поводом назвать данные большими является осознание того, что их новые свойства не дают возможности обрабатывать их на одном компьютере, но и этот факт не может быть признан четким критерием, ведь тут многое зависит от мощности и производительности компьютера.

Свойство разнообразия подчеркивает, что данные могут быть разного типа или разной структуры. Причем, в их массиве вполне допускаются пропуски. Форматы представления данных тоже могут различаться. Важно также осознавать, что для решения своей задачи специалист должен каким-то образом структурировать имеющиеся у него данные. Если бы речь шла о подконтрольной ему системе, он бы мог как-то настроить ее структуру и типологию, чтобы легче справляться с потоком информации. Но особенность больших данных состоит как раз в том, что системы, откуда они берутся, очень разные и в большинстве своем неподконтрольны субъекту, хотя он и имеет к ним доступ. Вот и получается, что, с одной стороны, есть великий соблазн использовать эти данные, но с другой – влиять на их структуру, скорость или объем практически невозможно.

Под скоростью тоже понимается не столько конкретное количество информации, переданной за единицу времени, сколько тот факт, что она поступает непрерывным потоком, не дающим времени для паузы, поэтому обрабатывать ее нужно «на лету». Технологии больших данных должны не только справляться с этой обработкой в режиме «онлайн», но и предоставлять пользователю возможность немедленно реагировать на изменения.

Как видим, понятие «большие данные» в некоторой степени субъективно и идет от осознания, что в рамках прежних подходов работать с данными невозможно. Для раскрытия их потенциала необходимы либо принципиально новые подходы к обработке

информации, либо новые сочетания уже существующих методов в расчете на синергетический эффект. Фактически количественное изменение в случае с большими данными становится качественным. Из отдельных информационных структур они превращаются в поток. Влиять на него нельзя, можно только использовать его. Если этого не сделать, поток данных просто пройдет мимо. Не более того.

Инструментарий для работы с большими данными как синергия существующих технологий

Для работы с Big data используются технологии, разработанные задолго до появления этого феномена. Среди них – искусственный интеллект и машинное обучение, средства визуализации данных, алгоритмы поиска, динамические структуры данных и многое другое. Так, машинное обучение и нейронные сети используются в защите информации [Перов, Пестунов, 2020; Монарев, Пестунов, 2018], сельском хозяйстве [Терещенко и др., 2022], физической географии [Павлова, 2017], прогнозировании физико-химических свойств веществ [Осипов, Криветченко, 2013] и т.д. Визуализация важна во многих сферах: маркетинге, журналистике, физике, математике и пр. Структуры данных и алгоритмы применяются при разработке самых обычных программных продуктов.

Однако после превращения данных в поток и обретения ими V-характеристик, с одной стороны, возникли новые требования к алгоритмам их обработки, с другой – они сами стали ресурсом и драйвером для тестирования, оптимизации и развития существующих технологий. Именно большие данные придали информационным технологиям новый импульс, предоставляя возможность проведения более масштабных экспериментов и ставя новые амбициозные задачи для исследователей и разработчиков.

Управление на основе данных как инновация в менеджменте

Совершенствование информационных технологий, в том числе технологий работы с большими данными, привело к возможности повысить эффективность использования информации при принятии управленческих решений [Попазова, Шихова, 2019; Маркова, Марков, 2019; Фиофанова, 2021]. Управление на основе данных позволяет существенно снизить значимость субъективных факторов, что повышает точность картины мира руководителей. Разумеется, последние всегда старались принимать объективные

решения с учетом своих целей и задач, но именно современные технологии работы с информацией открывают перспективы радикально повысить качество решений.

Типовой алгоритм управления на основе данных выглядит следующим образом.

1. *Фиксация проблемы.* На первом шаге руководитель или иное заинтересованное лицо формулирует проблему, которая его тревожит. Проблема может выражаться в низких показателях, плохих отзывах клиентов, дефиците бюджета и пр.

2. *Формулирование гипотез.* Специалисты вырабатывают одну или несколько гипотез относительно причин возникновения и возможных путей решения этой проблемы. Они подтвердятся или будут опровергнуты на следующих шагах.

3. *Определение датасета.* Последний представляет собой набор данных специального (часто табличного) формата, на основе которого проводится проверка гипотез различными способами, в том числе методами машинного обучения. Это предполагает формирование перечня и структуры данных, которые необходимо собрать.

4. *Сбор данных.* Этот этап предполагает не только непосредственное аккумулирование информации согласно структуре и источникам, определенным ранее. Проводятся также ее верификация, обработка и сохранение согласно сформированной структуре.

5. *Проверка гипотез.* На этом шаге определяется, какие из гипотез оказались верными и могут быть положены в основу управленческого решения.

6. *Формирование проекта и принятие управленческого решения.* Проект решения базируется на подтвержденных гипотезах. Решение фиксируется и воплощается в трансформации бизнес-процессов, перечне собираемых данных и т.д.

7. *Оценка эффективности принятого решения.* Строгих правил, регламентирующих алгоритм процедуры оценки, обычно не существует. Она может проводиться как на модельных, так и на реальных объектах.

8. *Повторная проверка гипотез.* На этом шаге может возникнуть потребность в коррекции правил принятия управленческих решений, оценка которых на предыдущем этапе показала их низкую эффективность.

Практика работы с большими данными свидетельствует, что при их анализе в ходе верификации гипотез нередко обнаружива-

ются неочевидные причинно-следственные связи. Закономерности и зависимости, выявленные на основе Big Data, зачастую невозможно предсказать заранее. Они могут оказаться неожиданными и контринтуитивными.

Большие данные требуют новых компетенций: кого следует нанять для получения пользы от их обработки?

Для руководителей, осознающих и желающих реализовать потенциал использования больших данных при принятии управленческих решений, актуален вопрос о том, какие специалисты способны это сделать. Детальный ответ на него зависит от имеющегося в распоряжении компании набора данных и их качества. Однако можно выделить три ключевые функциональные роли, не зависящие от конкретной ситуации: аналитик, инженер данных и BI-инженер. В небольшой компании или для начала работы с большими данными из трех таких специалистов вполне возможно сформировать структурное подразделение, способное приносить реальную пользу.

Аналитик. Его задача – выработка требований к структуре данных для проверки сформулированных гипотез. Аналитик взаимодействует с заказчиком (руководителем компании), который озвучивает проблемы, требующие решения, составляет перечень гипотез и ставит задачи по их проверке и оценке инженеру данных и BI-инженеру. В случае сложных или объемных задач их постановка может иметь вид технического задания.

Инженер данных. Это специалист технического плана, который подбирает и разрабатывает инструменты получения данных из корпоративных информационных систем и иных источников и занимается их переносом в единое хранилище (осуществляет ETL-процессы). В рамках своей деятельности инженер данных должен справляться со всевозможными нестыковками в имеющейся информации (ошибки, пропуски или искажения и пр.), порожденными как техническими причинами, так и влиянием человеческого фактора. Ведущему инженеру данных могут помогать менее квалифицированные специалисты, выполняющие задачи, связанные с рутинной обработкой информации и выполнением большого количества однообразных манипуляций (разметка карт, выгрузка информации, их перенос из одной системы в другую, перевод из неструктурированного вида в структурированный и т.п.).

BI-инженер настраивает и поддерживает систему визуализации отчетности, а также разрабатывает «дашборды» (сводные аналитические отчеты), реализуя требования, сформулированные заказчиком. Фактически такие дашборды являются главным результатом работы подразделения по работе с большими данными. В них в сжатом виде представлена информация для принятия управленческих решений. Дашборды могут быть как статическими, так и динамическими – обновляемыми в режиме реального времени.

Наличие в штате перечисленных специалистов не избавляет прочих сотрудников (от административно-управленческого персонала до технических специалистов) от необходимости владеть определенными цифровыми компетенциями. К примеру, бухгалтер уже не может выполнять свои обязанности, не имея навыков работы в специализированных программах и не ориентируясь в многочисленных государственных справочно-информационных системах. Специалист любого отдела должен уметь выполнять свои функции с использованием цифровых сервисов, многие из которых могут быть источниками данных, используемых в управлении и при принятии решений.

Получится ли упорядочить большие данные на основе всеобщего консенсуса?

Сложность работы с большими данными обусловлена сочетанием их неоднородной структуры и большого объема. Если устранить хотя бы один из этих факторов, задача обработки информации радикально упростится. На сокращение объема рассчитывать не приходится, поскольку развитие технологий способствует фиксации все большего количества данных. Что же касается структуры, здесь есть целый ряд аспектов, которые имеет смысл обсудить.

Для эффективной работы с информацией необходимо ее структурировать в определенном порядке. Правильно подобранная структура позволяет быстро выполнять различные операции с данными (добавление, поиск, преобразование и пр.), легко в них ориентироваться и принимать качественные решения на их основе. Именно поэтому многие из нас периодически наводят порядок в своих аккаунтах электронной почты, создавая правила обработки входящих писем, удаляя ненужные и сортируя оставшиеся. Поддержание такого порядка требует ресурсов (как минимум – сил и времени), и если говорить об эффективности

издержек, то польза от наведения порядка должна быть больше, чем затрачиваемые ресурсы.

Разнообразие (variety) и изменчивость (variability) больших данных обусловлены несогласованностью источников их генерации. Программные приложения, регистрирующие данные, нередко соответствуют разным стандартам, либо вовсе – неким ситуативным требованиям, что препятствует формированию стандартизованных потоков данных, с которыми можно было бы работать единообразно. Отметим, что такая ситуация многим выгодна. Коммерческие цели часто вынуждают компании выработать свои стандарты, чтобы отличаться от конкурентов. Именно этим была обусловлена, например, «война браузеров», активная фаза которой протекала в 1995–2001 гг. Формирование консенсуса для согласованной генерации данных с целью повышения эффективности их обработки – это настоящий вызов для человечества. Тем не менее пока на него нет ответа, субъект, желающий извлечь полезную информацию из больших данных, должен быть способен адаптироваться к изменению структуры, формата, типологии и других параметров поступающей информации. Скажем более: для использования всего потенциала, содержащегося в доступных данных, нельзя делать ставку только на те из них, структурой которых можно управлять. Требуемая информация может располагаться на неподконтрольных серверах и иметь неподконтрольную структуру.

Извлечение нужной информации из больших данных можно сравнить с переработкой мусора: есть понимание, что при соответствующей его переработке можно получить полезные ресурсы, но возможности повлиять на «формат» образующихся отходов крайне ограничены, особенно если приходится договариваться в рамках города, региона, страны и тем более – всего мира. При этом определенные успехи в данном направлении имеются, в частности, многие страны успешно реализуют отдельный сбор бытовых отходов.

Мы можем сохранять оптимизм в отношении выработки стандартов, определяющих структуру больших данных и некоего приведения их потока к такому виду, чтобы извлечь максимум пользы. Усилия по выработке консенсуса относительно структуры больших данных, безусловно, будут нарастать. Возможно, на помощь в этом деле придут технологии «блокчейн» [Пестунов, 2018].

Для примера рассмотрим ситуацию с развитием сети Интернет, которая появилась относительно недавно. До этого момента

существовали отдельные внутренние сети или автономные ЭВМ, не сообщающиеся друг с другом. Сегодня же практически любое устройство по умолчанию подключается к сети Интернет. Это стало возможным благодаря внедрению стандартов передачи данных. Несмотря на целый ряд сложностей в прошлом, теперь практически любой сайт открывается любым web-браузером.

Проблема защиты персональных данных и приватности приобретает новое значение

Возможность неявного извлечения информации из Big data по-новому ставит задачу защиты персональных данных. Закон 152-ФЗ² призван защищать пользователей лишь в том случае, когда речь идет о конкретной информации, предоставляемой оператору или публикуемой в открытых источниках. Согласие на использование персональных данных обязывает оператора корректно использовать и не разглашать только ту информацию, которая перечислена в документе. Если же явного предоставления персональной информации не происходило, то апеллировать к закону не получится. Поэтому граждане юридически не защищены при опосредованном извлечении информации с помощью технологий больших данных. Вероятна даже такая ситуация, когда пользователь явно не предоставлял никакой информации о себе, но третьи лица могут без его ведома выявить необходимые им сведения (в том числе те, которые он не желает разглашать).

В качестве примера приведем видеосъемку в общественных местах. Как правило, никакая персональная информация в таких местах не демонстрируется. Однако при анализе таких видеозаписей и/или их сопоставлении с другой открытой информацией (например, имеющейся в социальных сетях), можно многое узнать о круге общения, привычках, распорядке дня человека и т.д., т.е. ту информацию, которую тот вовсе не хотел бы афишировать. Подобная ситуация может возникнуть при анализе поисковых запросов пользователя в сети Интернет, хотя он при этом не предоставляет персональную информацию. Накопив определенный массив запросов, владельцы поисковой системы могут достоверно составить портрет пользователя по различным аспектам: состояние здоровья, интересы, образование и многое другое.

² Федеральный закон от 27.07.2006 № 152-ФЗ «О персональных данных».

Получается, что операторы больших данных формально не являются операторами данных персональных, поскольку далеко не всегда собирают их явно. Но с использованием соответствующих технологий могут получить информацию, которую пользователь разглашать не желает, не неся за это никакой ответственности. Это заставляет под другим углом взглянуть на проблему подписания информированного согласия [Andreotta et al., 2022].

Морально-этические аспекты асимметричного доступа к большим данным

Вопросам использования больших данных и прочих цифровых технологий для различного вида контроля уделяется существенное внимание в научной литературе [Прошунин, 2022; Абдулгалимов, 2021; Чхутиашвили, 2022; Смирнов, 2022; Coghlan et al., 2021]. При этом важны не только технические вопросы и обеспечение эффективности, но и морально-этические проблемы [Hoffmann, 2018; Mittelstadt, Floridi, 2016], одну из которых рассмотрим в данном разделе.

Технологии работы с большими данными и информационные технологии в целом предлагают широкий спектр полезных и удобных инструментов как отдельно взятому пользователю, так и крупным корпорациям. Однако есть риск, что с их помощью цифровые гиганты и государство станут контролировать всех и каждого в том или ином смысле. С одной стороны, действительно, поисковые системы, интернет-магазины и другие сервисы имеют доступ к огромным массивам пользовательских данных, которыми могут практически бесконтрольно пользоваться в своих целях, что вызывает законные опасения. Но ведь и сам пользователь получает доступ к большому объему своих и чужих данных, многие из которых находятся в открытом доступе (профили социальных сетей, данные всевозможных открытых реестров, доступ к информации программных продуктов через прикладные программные интерфейсы (application programming interface, API), автоматически записываемая история сообщений с инструментами работы с ней и пр.). Кроме того, ряд сервисов (например, социальные сети или интернет-магазины) предоставляет доступ пользователям и к чужим данным, что, очевидно, было бы невозможно без их предоставления и накопления.

В итоге можно сделать вывод о том, что крупные корпорации выступают не только в качестве «монстров», желающих

посредством сбора и анализа данных взять всех нас под контроль, но и в качестве провайдеров этих же инструментов для отдельных пользователей. Проблема здесь кроется в асимметричном доступе к данным. Корпорации аккумулируют информацию от всех пользователей, а пользователи – только свою и отчасти чужую, которую им на тех или иных условиях предоставляет сервис.

Но с этим ничего не поделаешь, ведь поддержание работы этих сервисов как раз и заключается в сборе и обработке данных. Другими словами, обозначенные опасения относительно усиления контроля или других негативных последствий, связанных с неправомерным использованием больших данных, – это вопрос доверия и добросовестности оператора и прочих субъектов, имеющих к ним доступ, а не призыв отказаться от этих технологий и полностью закрыть персональную информацию.

О социальных противоречиях, вызванных цифровизацией и феноменом больших данных

Несмотря на то, что новые возможности, которые открываются в эпоху больших данных, несомненно, имеют ряд преимуществ, тотальная цифровизация порождает и нежелательные социальные явления. Во всех отраслях наблюдается дефицит кадров, обладающих цифровыми навыками. Вообще, сегодня часто предъявляются специфические требования к профессиональным компетенциям и личностным характеристикам персонала (во всяком случае, отличающиеся от тех, что были еще 5–10 лет назад). Некоторые компании сталкиваются с сопротивлением персонала изменениям. До сих пор существует территориальное неравенство в части доступа к Интернету и современному оборудованию: разница в уровне цифровизации центра и периферии как на уровне страны в целом, так и отдельных ее регионов и районов очевидна.

Еще одна проблема связана с доступностью образовательных ресурсов. Очевидно, что тотальная цифровизация, одним из следствий которой и стало явление больших данных, влечет и требования к формированию новых цифровых компетенций персонала организаций, и необходимость его постоянного обучения и повышения квалификации. С одной стороны, доступ к образованию сейчас открыт для каждого: новые образовательные программы вузов на бюджетной и коммерческой основе, образовательные программы частных организаций, которые пропагандируются как знания «без воды», многочисленные открытые онлайн-курсы

с платными и бесплатными разделами, онлайн-школы и т.д. С другой стороны, воспользоваться этими возможностями зачастую бывает затруднительно: дорого, не хватает времени на качественное изучение материала после работы, не хватает базовых знаний, качество большинства образовательных продуктов оставляет желать лучшего. Все это ведет к еще одному социальному противоречию, которое особенно явно прослеживается в бюджетной сфере: уровень заработной платы не поспевает за ростом новых профессиональных требований.

Таким образом, серьезная задача управления кадровым потенциалом организаций в условиях цифровизации многократно усложняется. На первый план выходят задачи преодоления сопротивления персонала к изменениям и обеспечение повышения уровня его квалификации.

Заключение

Сформулируем ключевые тезисы, отражающие главные идеи, рассмотренные в статье.

1. Основная причина превращения массива обычных данных в «большие» состоит в том, что стало возможным автоматически регистрировать и затем сохранять практически любые события и изменения состояния объектов.

2. Зарегистрированные и сохраненные данные могут быть источником получения выгоды, для извлечения которой требуются технологии, позволяющие обрабатывать «сырые» данные и извлекать из них релевантную информацию.

3. Многие технологии, используемые для обработки Big data, не являются принципиально новыми. Скорее, можно говорить о том, что большие данные стали ресурсом для развития уже существующих технологий.

4. Большие данные очень плохо поддаются структуризации, конкретному субъекту повлиять на их структуру практически невозможно, и приходится работать с тем, что есть. От умения субъекта подстроиться под это обстоятельство зависит эффективность его работы и размер получаемой выгоды.

5. Вопрос о возможности установления глобального консенсуса о структуре данных до сих пор открыт; если он будет решен положительно, эффективность обработки больших данных совершит радикальный скачок.

6. Ключевая этическая проблема *Bis data* – это асимметричность доступа к информации. Хотя даже обычные пользователи используют сервисы на основе больших данных (поисковики, социальные сети и пр.), для владельцев этих сервисов доступность данных несоизмеримо выше.

7. Технологии больших данных по-новому ставят вопрос защиты персональной информации. Здесь уже недостаточно согласия пользователя на ее обработку, поскольку оператор способен опосредованно извлечь сведения, напрямую пользователем не предоставленные.

8. Три базовые роли, которые необходимы для внедрения в компании управления на основе данных, – это аналитик, инженер данных и *VI*-инженер.

9. Цифровизация и технологии больших данных породили новые социальные противоречия в обществе, которые требуют решения.

Литература

Абдулгалимов А.М. Государственный финансовый контроль в условиях развития цифровой экономики // Вестник Чеченского государственного университета им. А.А. Кадырова. 2021. № 3. С. 19–25.

Булгаков С.В. Большие данные в интеллектуальных транспортных системах // Наука и технологии железных дорог. 2022. Т. 6. № 2. С. 45–52.

Гвишиани А.Д., Добровольский М.Н., Дзеранов Б.В., Дзедобоев Б.А. Большие данные в геофизике и других науках о земле // Физика Земли. 2022. № 1. С. 3–34.

Засухина О.А., Еришов Е.В., Головатюков Л.К., Шитенков Г.А. Большие данные (big data) в области электроэнергетики // Вестник Ангарского государственного технического университета. 2022. № 16. С. 16–20.

Маркова В.Д., Марков А.А. Цифровизация, или управление на основе потока данных // Инновации. 2019. № 7 (249). С. 83–87.

Мильман Б.Л., Журкович И.К. Большие данные в современном химическом анализе // Журнал аналитической химии. 2020. Т. 75. № 4. С. 316–326.

Монарев В.А., Пестунов А.И. Эффективное обнаружение стеганографически скрытой информации посредством интегрального классификатора на основе сжатия данных // Прикладная дискретная математика. 2018. № 40. С. 59–71.

Осипов А.Л., Криветченко О.В. Компьютерная оценка нижнего концентрационного предела воспламенения // В мире научных открытий. 2013. № 10–1. С. 34–45.

Павлова А.И. Анализ методов интерполирования высот точек для создания цифровых моделей рельефа // Автометрия. 2017. № 2. С. 86–94.

Перов А.А., Пестунов А.И. О возможности применения сверточных нейронных сетей к построению универсальных атак на итеративные блочные шифры // Прикладная дискретная математика. 2020. № 49. С. 46–56.

Пестунов А.И. Криптовалюты и блокчейн: потенциальные применения в государстве и бизнесе // ЭКО. 2018. № 8. С. 78–92. DOI: 10.30680/ECO0131-7652-2018-8-78-92

Половинченко М.И., Елисеев В.С. Большие данные и их применение в агробизнесе // Мехатроника, автоматика и робототехника. 2021. № 7. С. 46–49.

Попазова О.А., Шихова Н.Н. Управление персоналом на основе анализа больших данных: риски и возможности // Известия Санкт-Петербургского государственного экономического университета. 2019. № 3. С. 110–115.

Прошунин М.М. Государственный цифровой финансовый контроль: правовая сущность // Российское правосудие. 2022. № 7. С. 98–104.

Рыльникова М.В., Макеев М.А., Кадочников М.В., Клебанов Д.А. Большие данные для оптимизации работы погрузочной техники и автотранспорта на горных работах // Известия Тульского государственного университета. Науки о Земле. 2022. № 4. С. 343–354.

Смирнов С.А. Наше бесчеловечное будущее или уловка трансгуманизма // Человек. 2022. № 1. С. 61–79.

Терещенко С.Н., Осипов А.Л., Моисеева Е.Д. Определение количества колосьев на изображениях пшеничных полей методами компьютерного зрения // Автометрия. 2022. № 3. С. 72–78.

Фюфанова О.А. Управление на основе больших данных в сфере образования // Государственная служба. 2021. Т. 23. № 3. С. 86–91.

Чхутиашвили Л.В. Государственный аудит и внутренний контроль в условиях цифровой экономики // Аудит. 2022. № 3. С. 9–11.

Andreotta A., Kirkham N., Rizzi M. (2022). AI, Big Data, and the future of consent. *AI & Society*. No. 37. Pp. 1715–1728. Springer

Coghlan S, Miller T, Paterson J. (2021). Good proctor or “Big Brother”? Ethics of online exam supervision technologies. *Philosophy & Technology*. No. 34. Pp. 1581–1606. Springer.

Floridi L. (2012). Big Data and Their Epistemological Challenge. *Philosophy & Technology*. No. 25. Pp. 435–437.

Hoffmann A. (2018). Making data valuable: political, economic, and conceptual bases of Big Data. *Philosophy & Technology*. No. 31. Pp. 209–212. Springer.

Mittelstadt B., Floridi L. (2016). The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. *Science and Engineering Ethics*. No. 22. Pp. 303–341. Springer.

Статья поступила 11.05.2023

Статья принята к публикации 09.07.2023

Для цитирования: Пестунов А.И., Гинтофт А.С., Криветченко О.В. Big data как феномен: причины и следствия появления больших данных // ЭКО. 2023. № 9. С. 137–154. DOI: 10.30680/ЕКО0131-7652-2023-9-137-154

Информация об авторах

Пестунов Андрей Игоревич (Новосибирск) – кандидат физико-математических наук, доцент. Новосибирский государственный университет экономики и управления.

E-mail: pestunov@gmail.com; ORCID: 0000-0002-4909-7953

Гинтофт Александр Сергеевич (Новосибирск) – Новосибирский государственный университет экономики и управления.

E-mail: gintoft.a.s@gmail.com

Криветченко Оксана Викторовна (Новосибирск) – Новосибирский государственный университет экономики и управления.

E-mail: kriv_ok@ngs.ru

Summary

A.I. Pestunov, A.S. Gintoft, O.V. Krivetchenko

Big Data as a Phenomenon: Causes and Consequences of the Emergence of Big Data

Abstract. The paper shows that many technologies, now called big data technologies, existed long before this term was introduced into mass use. The fundamental difference of the current situation is only in the scale of use and a certain synergetic effect resulting from the development of a number of initially loosely connected technologies. Alongside techno-economic problems, the paper considers humanitarian-ethical and social ones. The authors list professions that become relevant within the framework of big data technologies and the digital competencies demanded in this context. They also raise the problem of personal data protection in a novel way.

Keywords: *big data; data-based management; digital economy; digital competences; artificial intelligence; machine learning; neural network; Laplace's demon; determinism; personal data; ethics; morality; social contradictions of digitalization*

References

- Abdulgalimov, A.M. (2021). State financial control in the context of development digital economy. *Vestnik Kadyrov Chechen State University*. No 3. Pp. 19–25. (In Russ.).
- Andreotta, A., Kirkham, N., Rizzi, M. (2022). AI, Big Data, and the future of consent. *AI & Society*. No. 37. Pp. 1715–1728. Springer
- Bulgakov, S.V. (2022). Big data in intelligent transportation systems. *Science and technologies of railway*. Vol. 6. No. 2. Pp. 45–52. (In Russ.).
- Chkhutiashvili, L.V. (2022). State audit and internal control in the digital economy. *Audit*. No. 3. Pp. 9–11. (In Russ.).
- Coghlan, S., Miller, T., Paterson, J. (2021). Good proctor or “Big Brother”? Ethics of online exam supervision technologies. *Philosophy & Technology*. No. 34. Pp. 1581–1606. Springer.
- Fiofanova, O.A. (2021). Big data-driven management in education. *Public Administration*. Vol. 23. No. 3. Pp. 86–91. (In Russ.).
- Floridi, L. (2012). Big Data and Their Epistemological Challenge. *Philosophy & Technology*. No. 25. Pp. 435–437.
- Gvishiani, A.D., Dobrovolsky, M.N., Dzeranov, B.V., Dzeboev, B.A. (2022). Big data in geophysics and other earth sciences. *Izvestiya, Physics of the Solid Earth*. Vol. 58. No. 1. Pp. 1–29. (In Russ.).
- Hoffmann, A. (2018). Making data valuable: political, economic, and conceptual bases of Big Data. *Philosophy & Technology*. No. 31. Pp. 209–212. Springer
- Markova, V.D., Markov, A.A. (2019). Digitalization or data stream-based management потока данных. *Innovations*. No. 7. Pp. 83–87. (In Russ.).
- Milman, B.L., Zhurkovich, I.K. (2020). Big data in modern chemical analysis. *Journal of Analytical Chemistry*. T. 75. No. 4. C. 443–452. (In Russ.).
- Mittelstadt, B., Floridi, L. (2016). The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. *Science and Engineering Ethics*. No. 22. Pp. 303–341. Springer.
- Monarev, V.A., Pestunov, A.I. (2018). Efficient steganography detection by means of compression-based integral classifier. *Prikl. Diskr. Mat.* No. 40. Pp. 59–71. (In Russ.).

Osipov, A.L., Krivetchenko, O.V. (2013). Computer assessment of lower flammability limit of chemicals. *In the World of Scientific Discoveries*. No. 10–1. Pp. 34–45. (In Russ.).

Pavlova, A.I. (2017). Analysis of elevation interpolation methods for creating digital elevation models. *Optoelectronics, Instrumentation and Data Processing*. Vol. 53. No. 2. Pp. 171–177. (In Russ.).

Perov, A.A., Pestunov, A.I. (2020). On possibility of using convolutional neural networks for creating universal attacks on iterative block ciphers. *Prikl. Diskr. Mat.* No. 49. Pp. 46–56. (In Russ.).

Pestunov, A.I. (2018). «Blockchain» distributed secure ledger and cryptocurrencies: potential using in business and government. *ECO*. No. 8. Pp. 78–92. (In Russ.). DOI: 10.30680/ECO0131-7652-2018-8-78-92

Polovinchenko, M.I., Eliseev, V.S. (2021). Big data and its application in agribusiness. *Mehatronics, automatics and robototechnics*. No. 7. Pp. 46–49. (In Russ.).

Popazova, O.A., Shikhova, N.N. (2019). Personnel management based on big data analytics: risks and opportunities. *Izvestia St. Petersburg State University of Economics*. No. 3. Pp. 110–115. (In Russ.).

Proshunin, M.M. (2022). The state digital financial control: legal nature. *Russian Justice*. No. 7. Pp. 98–104. (In Russ.).

Rylnikova, M.V., Makeev, M.A., Kadochnikov, M.V., Klebanov, D.A. (2022). Leverage of big data to optimize the operation of loading equipment and vehicles in surface mining. *Izvestia Tula State University. Earth Sciences*. No. 4. Pp. 343–354. (In Russ.).

Smirnov, S.A. (2022). Our inhuman Future or Tricks of Transhumanism. *Chelovek*. Vol. 33. No. 1. Pp. 61–79. (In Russ.).

Tereshchenko, S.N., Osipov, A.L., Moiseeva, E.D. (2022). Determination of the number of ears in images of wheat fields by methods computer vision. *Optoelectronics, Instrumentation and Data Processing*. Vol. 58. No. 2. Pp. 72–78. (In Russ.).

Zasukhina, O.A., Ershov, E.V., Golovatyukov, L.K., Shitenkov, G.A. (2022). Big data in the electric power industry section. *Vestnik Angarsk State Technical University*. No. 16. Pp. 16–20. (In Russ.).

For citation: Pestunov, A.I., Gintoft, A.S., Krivetchenko, O.V. (2023). Big Data as a Phenomenon: Causes and Consequences of the Emergence of Big Data. *ECO*. No. 9. Pp. 137–154. (In Russ.). DOI: 10.30680/ECO0131-7652-2023-9-137-154

Information about the authors

Pestunov, Andrey Igorevich (Novosibirsk) – Candidate of Physical and Mathematical Sciences, Associate Professor. Novosibirsk State University of Economics and Management.

E-mail: pestunov@gmail.com; ORCID: 0000–0002–4909–7953

Gintoft, Alexander Sergeevich (Novosibirsk) – Novosibirsk State University of Economics and Management.

E-mail: gintoft.a.s@gmail.com

Krivetchenko, Oksana Viktorovna (Novosibirsk) – Novosibirsk State University of Economics and Management.

E-mail: kriv_ok@ngs.ru