УДК 1:001

БОЛЬШИЕ ДАННЫЕ: ОТ БРАГЕ – К НЬЮТОНУ

Д.А. Новиков

Проанализированы и структурированы вызовы, которые технологии оперирования большими данными ставят перед разработчиками соответствующих технических средств, специалистами в области прикладной математики и искусственного интеллекта и учеными-предметниками из различных прикладных областей.

Ключевые слова: большие данные, анализ данных, большое управление.

Природа устроена просто. Надо лишь уметь находить надежные средства раскрытия этой осложненной подробностями простоты.

Э. Резерфорд

ВВЕДЕНИЕ

Термин «большие данные», обозначающий неструктурированные данные, объем которых превосходит существующие возможности оперирования ими в требуемые сроки, появился совсем недавно [1], тем не менее, он уже стал сверхпопулярным (запрос в Google возвращает десятки миллионов ссылок) как среди специалистов в IT-сфере, так и среди ученых, бизнес-аналитиков и многих других. Какие возможности и опасности несут большие данные? Какие вызовы они формулируют и какие проблемы ставят перед учеными, специалистами в различных предметных областях, системой образования?

1. БОЛЬШИЕ ДАННЫЕ. БОЛЬШАЯ АНАЛИТИКА. БОЛЬШАЯ ВИЗУАЛИЗАЦИЯ

Большие данные (Big Data, первое упоминание, — по-видимому, в специальном выпуске журнала «Nature» [1]) в информационных технологиях — направление в науке и практике, связанное с разработкой и применением методов и средств оперирования большими объемами неструктурированных данных.

Оперирование Big Data включает их[1]:

— сбор (получение);

— передачу;

— хранение (включая запись и извлечение);

— обработку (преобразование, моделирование, вычисления и анализ данных);

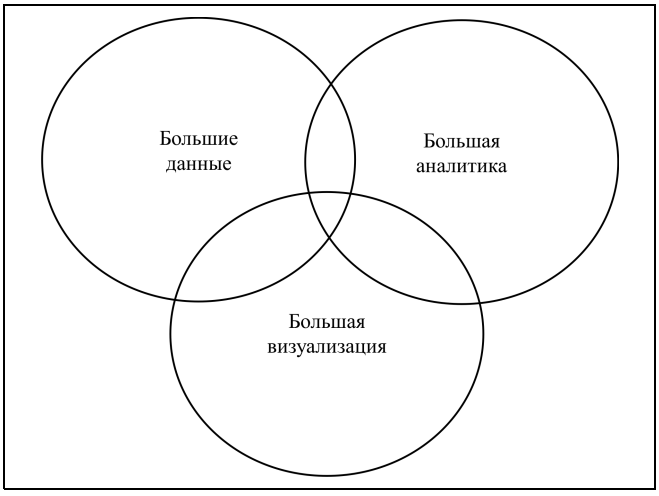
— использование (включая визуализацию) в практической, научной, образовательной и других видах человеческой деятельности.

Иногда «большими данными» в узком смысле называют только технологии сбора, передачи и хранения Big Data. Тогда обработку больших данных, включая построение и анализ моделей на их основе, называют *большой аналитикой* (в том числе и большие вычисления), а визуализацию (учитывающую когнитивные возможности пользователя) соответствующих результатов — *большой визуализацией* (рис. 1).

Универсальный цикл оперирования большими (да и, вообще, любыми) данными приведен на рис. 2. Ключевые элементы в данном цикле — *объект* и *субъект* («потребитель»), которому требуются знания о состоянии (и закономерностях его изменения) первого. Но между *данными*, собираемыми об объекте, и *знаниями*, необходимыми субъекту, иногда лежит целая «пропасть». Первичные данные должны быть предобработаны — превращены в более или менее структурированную *информацию*, из которой в зависимости от задачи, стоящей перед субъектом, должны быть извлечены требуемые знания.

Эти знания, в частности, могут быть использованы субъектом для *управления* объектом — осуществления целенаправленных воздействий, обеспечивающих требуемое его поведение. В частном случае (при неодушевленном субъекте) управление может быть автоматическим. Наверное, скоро

[1] Иногда говорят о конструкции «4D» — выявление (Discovery), отбор (Discrimination), переработка (Distillation), доведение в нужном представлении (Delivery/Dissemination).



#Рис. 1. «Большая триада»[2]

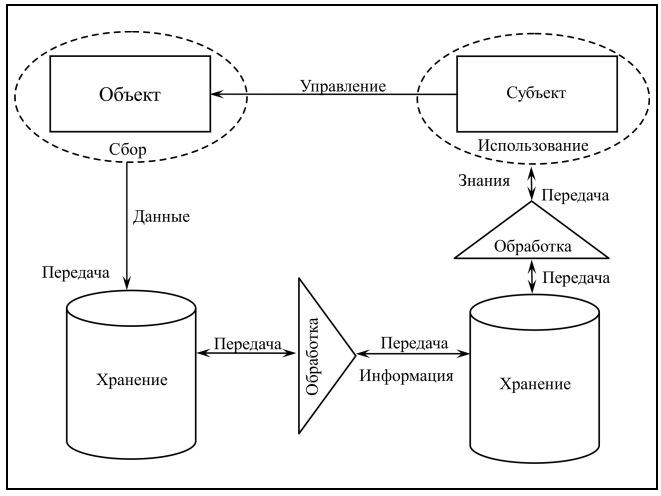


Рис. 2. Универсальный цикл оперирования Big Data

в обиход войдет термин «*большое управление*[3]» (Big Control) как управление на основе больших данных, большой аналитики и, быть может, большой визуализации[4].

Качественный анализ огромного потока текущих публикаций по Big Data позволяет сделать субъективную (авторскую) экспертную оценку текущего распределения внимания исследователей и разработчиков (но не пользователей!) к проблемам оперирования Big Data, представленную на рис. 3.

Другими словами, подавляющее большинство усилий в области Big Data направлено на разработку технологий сбора, передачи, хранения и предобработки больших данных, в то время как большой аналитике и визуализации уделяется гораздо меньшее внимание.

2. ЦИВИЛИЗАЦИОННЫЕ ПРОБЛЕМЫ

Можно ли считать сложившееся состояние дел (см. рис. 3) нормальным? С одной стороны — да. Ведь эволюционное развитие технологий шло именно этим путем, да и для того, чтобы анализировать и визуализировать данные, их надо сначала собрать и сохранить (естественно, таким образом, чтобы иметь возможность быстрого доступа и обработки). С другой стороны, существующий «перекос» вызван отчасти тем, что сегодня человечество, хотя и сознает, что, наверное, любые данные небесполезны, но не до конца понимает, что делать и как использовать их нарастающую лавину.

Проблема эта не нова, так как за последнее время возник целый класс подобных проблем, и носят они цивилизационный характер. Условно их можно назвать *проблемами опережающего развития технологий*: если рассмотреть соотношение между наукой, технологиями и практикой (рис. 4), то в

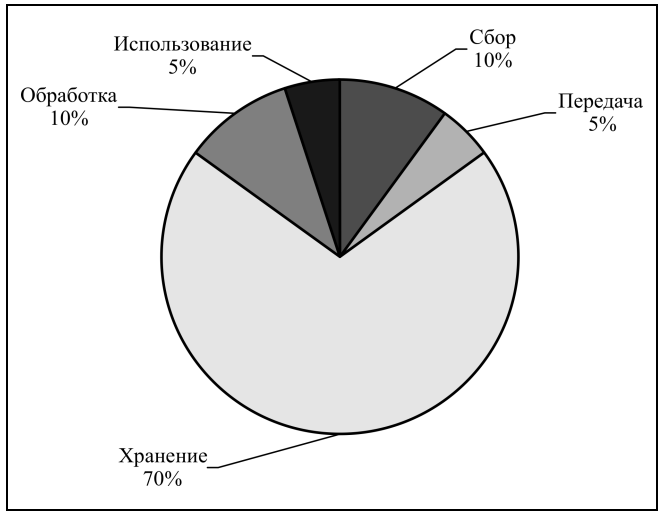


Рис. 3. Текущее распределение внимания исследователей и разработчиков к проблемам оперирования Big Data

[2] Еще одну сверхмодную триаду — большие данные, высокопроизводительные вычисления и облачные технологии — обсуждать мы не будем.

[3] Справедливости ради отметим, что специалисты по теории управления в последние полтора десятилетия все чаще говорят о совместном решении задач управления, вычислений и связи — так называемая проблема C^3 (Control, Computation, Communication) — решения задач синтеза управляющих воздействий в реальном времени с учетом задержек в каналах связи и временных затрат на обработку информации (включая вычисления). Кроме того, существует устойчивое словосочетание «управление большими системами» (Large-scale Systems Control), однако большие данные могут порождаться и «маленькими» в этом смысле системами.

[4] Возможна и другая трактовка термина «большое управление» — как управление процессами оперирования Big Data, что представляет собой самостоятельную и нетривиальную проблему.

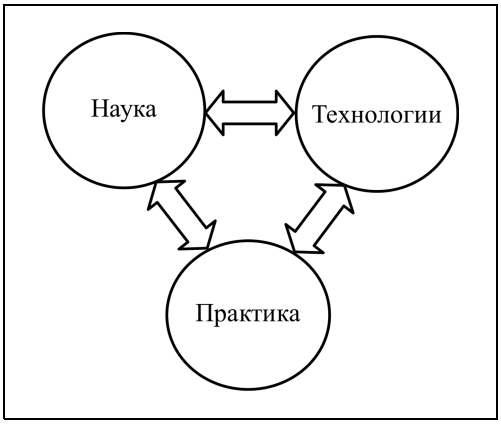


Рис. 4. Наука, технологии, практика

различные периоды развития человечества иногда наука инициировала развитие и последующее внедрение тех или иных технологий, а иногда последовательность была (и является в наше время!) «обратной».

Действительно, обратимся к истории. Сто лет, начиная примерно с середины XIX в., наблюдались триумфальное развитие науки и основывающийся на нем взрывной рост технологий (наука опережала и формировала технологии, которые массово внедрялись в практику; отметим, что до этого, как правило, опережающий запрос к науке формировала практика) — электричество, связь, атомная энергетика, электроника и т. д., причем технологий, понятных и доступных обывателю[5]. Затем, где-то с последней трети XX в. ситуация стала изменяться — уже накопленных фундаментальной наукой результатов хватало, чтобы обеспечить рост технологий (технологии начали в чем-то «опережать» науку), что привело к снижению «спроса на науку» со стороны последних (исключение представляют собой, наверное, лишь живые системы).

Но технологии продолжают развиваться, причем все более быстрыми темпами. Более того, например, в последнее десятилетие скорость развития информационно-коммуникационных технологий (ИКТ) стала опережать и практику[6], в том числе способности человечества по осознанию новых технологических возможностей, перспектив развития и соответствующих угроз[7]. Другими словами, на пороге третьего тысячелетия произошел определенный перелом — до сих пор человечество целенаправленно развивало науку и технологии под свои потребности, сейчас же технологии все больше опережающе навязывают направления, ограничения и условия развития, причем как на уровне отдельной личности, так и на уровне государств и человечества в целом[8]. Именно этот эффект мы и называем «опережающим развитием технологий». Осознать соответствующие цивилизационные проблемы и научиться реагировать на них еще предстоит.

То же самое происходит с большими данными — человечество обрело технологические возможности накопления гигантских объемов данных, но не готово их обрабатывать и использовать, причем основная проблема заключается не в непонимании того, *как* их обрабатывать, а в том, *зачем* это делать. Для того чтобы понять, какие вызовы стоят перед учеными и инженерами, обсудим кратко, какие данные являются «большими», где они возникают, как сегодня применяются, как повысить эффективность их использования в будущем и чем в этом может помочь наука.

3. ИСТОЧНИКИ И «ПОТРЕБИТЕЛИ» BIG DATA

Среди них такие крупные группы, как:

— наука — астрономия и астрофизика, метеорология, ядерная физика, физика высоких энергий, геоинформационные и навигационные системы, дистанционное зондирование Земли, геология и геофизика, аэродинамика и гидродинамика, генетика, биохимия и биология и др.

[5] Забавно, но героями фантастической литературы о технологических рывках в развитии, будь то благодаря изоляции (например, Сайрес Смит в «Таинственном острове» Ж. Верна) или переносу в прошлое (например, герой М. Твена из «Янки из Коннектикута при дворе короля Артура»), выступают, как правило, инженеры или просто обыватели, но не ученые.

[6] В качестве позитива, необходимо признать, что развитие ИКТ существенно стимулировало развитие прикладной математики, в том числе таких разделов «сетевой математики», как случайные графы, графы большой размерности, сетевые игры и многие другие — см. обзоры в работах [2—8].

[7] Отдельные проявления этого явления встречались и раньше — например, об этической ответственности ученых задумывались и А. Нобель, и участники Манхэттенского проекта и др.

[8] Проблемы информационной безопасности уже стали всем привычны. Но сейчас пора задуматься не только о технологической (кибербезопасности), но в более широком смысле — о социальной, экономической и другой безопасности информационных технологий — защищенности пользователей ИКТ, их групп и общества в целом от информационных воздействий (ярким примером являются социальные медиа, в том числе онлайновые социальные сети [9]). А поскольку все существенные решения (начиная с малого предприятия и заканчивая страной) принимаются на основе информации, которая откуда-то поступает, как-то (и не всегда известным лицу, принимающему решение, образом) передается и обрабатывается, то придется признать и важность социально-экономической безопасности ИКТ — защищенности личности, экономики, общества и государства от последствий решений, принимаемых с использованием современных ИКТ (в том числе, систем поддержки принятия экономических, военных, политических и других решений).

— Интернет (в широком смысле) и другие телекоммуникационные системы;

— бизнес, торговля и финансы, а также маркетинг и реклама (включая трейдинг, таргетирование и рекомендательные системы, CRM-системы, RFID — радиочастотные идентификаторы, все чаще используемые в торговле, транспорте и логистике и др.);

— мониторинг (гео-, био-, эко-; космический, авиа- и др.);

— безопасность (системы военного назначения, антитеррористическая деятельность и др.);

— электроэнергетика (включая атомную), Smart Grid;

— медицина;

— госуслуги и госуправление;

— производство и транспорт (объекты, узлы и агрегаты, системы управления и др.).

Многочисленные примеры[9] приложений Big Data в этих областях можно найти в научно-популярной (иногда даже в «глянцевой») литературе, свободно доступной в Интернете (повторять эти типовые примеры и «пугать» читателя «зетта-» и «йоттабайтами» мы не будем).

Современный уровень автоматизации практически всех перечисленных отраслей таков, что в них большие данные, де факто, автоматически генерируются. Поэтому все чаще задаются вопросом — сколько потоковых данных мы «теряем» (из-за того, что не можем или не успеваем их сохранить или обработать)? Вопрос этот корректен для инженера по ИКТ, но не для ученого и, тем более, не для пользователя результатов обработки Big Data — они бы спросили соответственно: «что мы потеряли существенного» и «что изменилось бы для нас, если бы мы успели все собрать и обработать».

4. КАКИЕ ДАННЫЕ ЯВЛЯЮТСЯ «БОЛЬШИМИ». ВЫЗОВЫ НАУКЕ

Традиционно, большими считаются, как минимум, данные, объем которых превосходит существующие возможности оперирования ими в требуемые сроки. Такое определение несколько «лукаво» — данные, считающиеся «большими» сегодня, перестанут быть таковыми завтра с развитием методов и средств работы с ними. Данные, казавшиеся «большими» несколько столетий или даже десятилетий назад (в отсутствии возможности их автоматической обработки), сегодня легко обрабатываются на бытовом компьютере. Соревнование между вычислительными потребностями (гипотетическими) человечества и соответствующими технологическими возможностями существует давно, и, естественно, потребности всегда опережали и будут опережать возможности. И несоответствие между ними служит колоссальным стимулом развития науки — приходится искать более простые (но адекватные) модели, придумывать более эффективные алгоритмы и т. д.

Иногда в определение Big Data добавляют такие их свойства, как 5V — объем (Volume), скорость (Velocity), разнообразие (Variety), достоверность (Veracity) и обоснованность (Validity); или говорят, что от большого объема обычных данных большие данные отличаются наличием большого потока (здесь учитывается и объем, и скорость — объем в единицу времени) неструктурированных[10] данных.

Неструктурированность (в широком понимании) Big Data (текст, видео, аудио, структура коммуникаций и т. п.), действительно, является их характерной чертой и вызовом для прикладной математики, лингвистики, когнитивных наук и искусственного интеллекта — разработка технологий обработки[11] в реальном времени, в том числе с возможностью выявления «скрытой» информации, больших потоков текстовой, аудио-, видео- и другой информации составляется мэйнстрим приложений перечисленных научных областей[12] к ИКТ.

Тем самым, мы наблюдаем прямой (и явный) запрос от технологий к науке. Второй (и столь же явный) «запрос» заключается в адаптации к анализу больших данных традиционных методов статистического анализа, оптимизации и т. п. Более того, помимо адаптации, необходима разработка новых методов, учитывающих специфику Big Data — сейчас модно рекламировать средства аналитики (как правило, бизнес-аналитики) для больших данных — см., например, критический анализ в рабо-

[9] Основная идея использования Big Data заключается в попытке выявить «скрытые закономерности» — найти ответы на нетривиальные вопросы, например: прогноз эпидемий по информации из соцсетей или о продажах в аптеках; задачи диагностики (медицинской и технологической); удержание клиентов благодаря анализу поведения покупателей в магазине по перемещению в пространстве RFID-меток товаров и пр.

[10] Неструктурированность данных может порождаться также их пропусками и/или разномасштабностью (в пространстве и времени — так называемых multi-scale systems) анализируемых явлений и процессов.

[11] Эти технологии, в первую очередь, должны предусматривать агрегирование данных (например, фиксация изменений в технологических данных или хранение агрегированных показателей), ведь не всегда нужно использовать все данные (особенно, если они «однородны»).

[12] Математика может хорошо работать со структурированными данными, поэтому преобразование неструктурированных данных в структурированные — отдельная важная задача.

те [10], но список этих средств почти совпадает с хрестоматийным набором статистических инструментов (и даже *уже* этого набора, так как не все методы применимы в условиях большой размерности). То же самое относится и к:

— методам машинного обучения (нейронные сети, байесовы сети, нечеткий логический вывод и т. п.);

— оптимизационным задачам большой размерности (как «альтернатива», помимо ставших привычными технологий параллельных вычислений, активно развивается распределенная оптимизация, l_1-оптимизация — см., например, обзоры и результаты в работах [11—15]);

— методам дискретной оптимизации (здесь «альтернативой» служит применение мультиагентных программных систем — см., например, [16]) и др.

Общее для перечисленных «запросов» технологий к науке заключается в том, что речь идет об адаптации или небольшой модификации известных, уже хорошо зарекомендовавших себя методов. Нужно понимать, что автоматическое построение (с помощью традиционного аппарата[13]) модели по сырым данным — в общем случае представляет собой не более чем модную иллюзию[14] — мы придумаем алгоритмы, «напустим» их на большие массивы неструктурированной (и в большинстве случаев нерелевантной) информации и благодаря этому будем принимать более эффективные решения. Подобные заблуждения уже встречались в истории науки — на ранних стадиях развития кибернетики и искусственного интеллекта[15] — и, приведя к множеству разочарований, очень сильно затормозили развитие этих научных направлений. Чудес на свете не бывает: как правило, для получения качественно новых выводов нужна новая модель, новая парадигма (см. работы по методологии науки [17, 18]).

5. ОБЩИЙ ВЫЗОВ

Сложность окружающего нас мира растет не так быстро, как возможности фиксации («измерения») и хранения данных, которые, похоже, опередили возможности человечества по осознанию возможности и целесообразности их использования, т. е., мы «захлебываемся» в данных и судорожно пытаемся придумать, что с ними можно делать.

Но на эту ситуацию можно посмотреть и с другой стороны: основной тезис заключается в следующем — получить *большие* (сколь угодно большого мыслимого объема) данные можно и достаточно просто (лежащие на поверхности примеры нам дают комбинаторная оптимизация, нелинейная динамика или термодинамика — см. далее), нужно понимать, что с ними делать (Природе нужно задавать правильные вопросы). Более того, можно придумать сколь угодно сложную модель, использующую Big Data, а затем пытаться достичь в ее рамках все более высокой точности. Проблема в том, получим ли мы при этом, кроме массы новых проблем[16], качественно новые результаты. Математикам и физикам давно известно, что увеличение размерности модели и ее «усложнение» (стремление учесть все больше факторов и связей между ними) далеко не всегда ведет к адекватном росту «качества» результатов моделирования, а иногда и вовсе приводит к абсурду[17].

Рассмотрим ряд примеров.

Пример 1. В книге нобелевского лауреата Г. Саймона [19] рассматривается следующий пример. Предположим, что мы наблюдаем за тем, как муравей движется по песку. Целью муравья может быть стремление минимизировать затраты своей энергии по перемещению из одной точки в другую, поэтому он огибает горки песка, иногда поворачивает назад и т. д. Если мы наблюдаем только проекцию на горизонтальную плоскость траектории муравья, а рельеф, по которому двигался муравей, неизвестен, то объяснить поведение муравья (сложную, петляющую траекторию) довольно непросто. Г. Саймон делает вывод, что наблюдаемое разнообразие и сложность поведения людей объясняются не сложностью принципов принятия ими решений, которые сами по себе просты, а разнообразием ситуаций, в которых принимаются решения. С этим мнением вполне можно согласиться. Действительно, нетривиальные результаты может давать как сложная модель на простых входных данных, так и простая модель на сложных данных. Желательно (в иде-

[13] Обычно ситуация дополнительно обременена предшествующим опытом исследователя/разработчика и традициями его научной школы — успешное решение некоторой одной задачи приводит к формированию вполне естественного убеждения, что этими же методами (ими и только ими!) можно решить все остальные нерешенные задачи.

[14] Хотя в некоторых случаях увеличение объема данных может (при правильной обработке) дать дополнительную информацию.

[15] Кибернетическая система не может продемонстрировать поведения, отличного от являющегося результатом заложенных в нее алгоритмов (которые могут быть «стохастическими», «недетерминированными» и т. д.), несмотря на кажущуюся генерацию новых знаний или проявления качественно нового («неожиданного») поведения, особенно при взаимодействии нескольких и, тем более, значительного числа элементов.

[16] О проблемах адекватности моделей и устойчивости результатов моделирования мы, осознавая их важность, пока забудем.

[17] Не говоря уже о ситуациях, когда в рамках существующих научных парадигм принципиально невозможно моделирование поведения системы на достаточно большом горизонте времени (примером может служить «точное» прогнозирование погоды).

але) уметь получать нетривиальные результаты в рамках простой модели, правильно выбрав для нее релевантные простые входные данные (недаром математики говорят, что простота — критерий истины).

Пример 2. Предположим, что в руки ученых, например, XVIII века волшебным образом попали современный ноутбук и компьютерный томограф (с инструкциями по эксплуатации). Сделав томограмму ноутбука и сохранив ее в последнем, они, вряд ли, анализируя полученные данные (очень подробные и многочисленные) о внутреннем «физическом» устройстве ноутбука, поняли бы хоть что-то о том, как он работает: т. е., правильная парадигма, правильная концептуальная модель являются необходимым (но, к сожалению, не достаточным) условием успеха.

Пример 3. Если взять любую NP-трудную задачу комбинаторной оптимизации [20], например — задачу коммивояжера, то в ней существует порядка $n!$ подлежащих анализу (в общем случае для поиска точного решения) вариантов, что уже для $n \sim 100$ превышает вычислительные возможности человечества. Это свойство NP-трудных задач известно давно - уже не одно десятилетие оно стимулирует специалистов на разработку методов поиска приближенных решений (с оценкой гарантированной «точности») за разумное время. Аналогичным примером могут служить модели нелинейной динамики — результатами «наблюдений» за динамическим хаосом, демонстрируемым даже достаточно простой (небольшой размерности) нелинейной динамической системой, можно занять память всех компьютеров Земли, но новых знаний эти данные содержать не будут.

Пример 4. Хрестоматийным примером из истории физики служит открытие закона всемирного тяготения. Тихо Браге (1546—1601) в течение двух десятилетий регулярно наблюдал за движением планет Солнечной системы. Его записи представляют собой Большие (по тем временам) данные. Иоганн Кеплер (1571—1630), на основе данных Браге, сформулировал свои эмпирические (!) законы движения планет.

Три закона Кеплера агрегировали информацию Браге, и движение каждой конкретной планеты могло быть рассчитано по ним (а не по многотомным записям Браге) с высокой точностью. Другими словами, Браге научился описывать[18] движение планет; Кеплер — описывать и предсказывать это движение. Но законы Кеплера ничего не говорят о том, *почему* планеты движутся в соответствии с этими законами. Ответ на этот вопрос (т. е. объяснение) дал закон всемирного тяготения И. Ньютона (1643—1727). Вывести законы Кеплера, наверное, мог бы любой современный компьютер (если в него заложить адекватные задаче алгоритмы и исходные данные), а сформулировать закон обратных квадратов — ни один (если в него не заложить соответствующую модель взаимодействия масс). Но, законы Кеплера являются «следствиями» закона тяготения (и могут быть выведены из него), как, в свою очередь, результаты Браге могут быть получены из законов Кеплера. Таким образом, закон всемирного тяготения «сделал ненужными» (точнее — гносеологически избыточными) и законы Кеплера, и большие данные Браге.

Сегодняшний опыт оперирования большими данными пока свидетельствует, что мы в большинстве случаев находимся на уровне Браге, предпринимая титанические попытки достичь уровня Кеплера. Но качественный скачок возможен только тогда, когда появляются обобщения (уровня Ньютона), радикально упрощающие ситуацию. Повторим: искусство заключается в том, чтобы «задавать Природе правильные вопросы».

Пример 5. Вторым хрестоматийным примером из истории физики служит создание молекулярно-кинетической теории, который свидетельствует, что породить большой поток данных — не проблема, вопрос в том, что мы хотим с этими данными делать и на какие вопросы отвечать.

Рассмотрим следующий мысленный эксперимент — задачу детального описания поведения идеального газа. Предположим, что при нормальных условиях находится один кубический метр воздуха. В нем содержится примерно 10^{25} молекул, движение каждой из которых и их соударения исчерпывающе (в рамках модели идеального газа) описываются кинематикой и динамикой, т. е. с точки зрения физики никаких принципиальных проблем описания их движения и взаимодействия не существует. Каждая молекула за одну секунду испытывает порядка 10^9 столкновений с другими молекулами. Описание поведения такой системы (координаты и скорости всех молекул) в реальном времени породит поток данных порядка 10^{35} байт/с.

Такой поток данных превосходит технологические возможности человечества даже сегодня! Наверное поэтому, еще в середине XIX в. при создании молекулярно-кинетической теории газов физики поняли бессмысленность детального анализа (при четком осознании его принципиальной возможности) и перешли к макроописанию в терминах агрегированных характеристик — температуры, объема, давления, а в дальнейшем в рамках статистической физики — к описанию в терминах вероятностных распределений. А ведь, если бы тогда смогли все «посчитать», остались бы мы без статистической физики! ♦

Завершая рассмотрение примеров, отметим, что большие данные по своему источнику можно условно разделять на *естественные* и *искусственные*. В первом случае данные порождает некоторый существующий независимо от нас объект, а мы (как «исследователи») решаем, что и сколько «измерять» и т. д. (см. примеры 1, 2 и 4). Во втором случае источником данных служит модель, которая может порождать большие данные (см. примеры 3 и 5), при этом сложность (поток данных) отчасти управляема и определяется в процессе моделирования.

[18] Напомним основные функции научного познания (в том числе и моделирования) [18]: описание (феноменологическая функция) — объяснение — прогноз (прогностическая функция) — управление (нормативная функция).

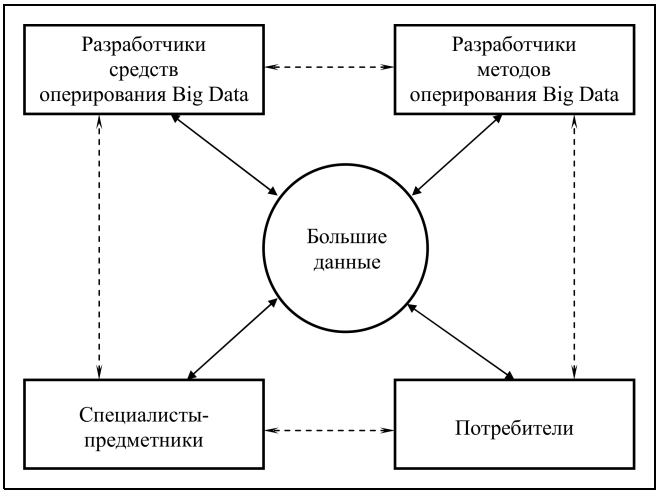


Рис. 5. Субъекты, оперирующие большими данными

«Рецепты». Можно выделить четыре большие группы субъектов (рис. 5), оперирующих (явно или косвенно) большими данными в своей профессиональной (научной и/или практической) деятельности:

— разработчики средств оперирования Big Data (производители соответствующего программного и аппаратного обеспечения, а также его продавцы, консультанты, интеграторы и др.);

— разработчики методов оперирования Big Data (специалисты по прикладной математике и компьютерным наукам);

— специалисты-предметники — ученые, исследующие реальные объекты (или их модели), служащие источниками больших данных;

— потребители, использующие или собирающиеся использовать результаты анализа больших данных в своей практической деятельности.

Каждый из представителей перечисленных групп взаимодействует с другими (см. штриховой контур на рис. 5) — нормативное (к которому надо стремиться) разделение «зон ответственности» приведено на рис. 6, где толщина стрелок условно отражает степень вовлеченности).

Не претендуя на конструктивность, даже исходя из здравого смысла, можно сформулировать следующие общие «рецепты» для перечисленных групп субъектов.

Для разработчиков средств оперирования Big Data: продавать решения в области Big Data (в том числе и аналитические) станет все труднее, если их не пополнять новыми адекватными математическими методами и не предусматривать возможность работы потребителя в тесном взаимодействии с разработчиками методов и специалистами-предметниками.

Для собратьев-математиков: актуален запрос на адаптацию известных и развитие новых (в первую очередь — обладающих линейной сложностью!) методов обработки больших потоков неструктурированных данных, которые представляют собой хороший полигон тестирования новых моделей, методов и алгоритмов (желательно за счет разработчиков и/или потребителей).

Для специалистов-предметников: технологии Big Data дают новые возможности получения и хранения огромных массивов «экспериментальной» информации, постановки так называемых вычислительных экспериментов, а развиваемые методы прикладной математики дают возможность системной генерации и быстрой верификации гипотез (выявления скрытых закономерностей).

Для потребителей: дорогие технологии сбора и хранения Big Data вряд ли дадут эффект без привлечения специалистов по методам и по предмету

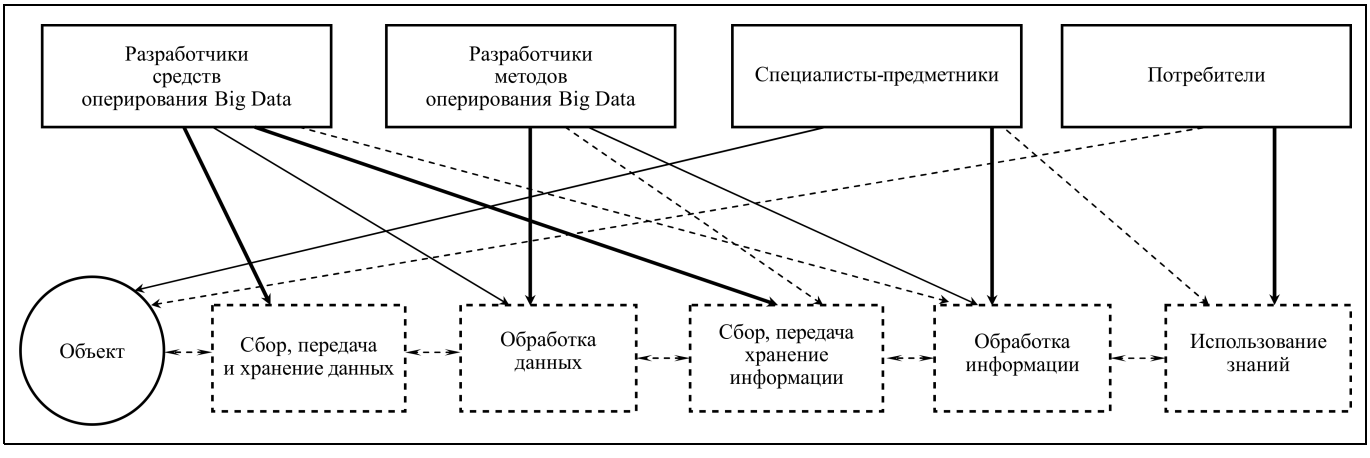


Рис. 6. Разделение «зон ответственности»

(при обязательном четком понимании, на какие вопросы потребитель хочет получить ответы от Big Data[19]).

Некоторые опасности. Помимо упомянутых необходимости поиска адекватных простых моделей и настораживающего тренда опережающего развития технологий, можно предположить будущую актуальность следующих проблем (их список неструктурирован и открыт).

- *Информационная безопасность Big Data*. Здесь потребуется и адаптация известных, и разработка принципиально новых методов и средств.
- *Энергетическая эффективность Big Data*. Уже сейчас центры обработки данных представляют собой существенный класс потребителей электроэнергии. Чем больше данных мы хотим обрабатывать, тем больше потребуется энергии.
- *Принцип дополнительности* давно известен в физике — измерение изменяет состояние системы. А как обстоит дело в социальных системах, элементы которых (люди) активны — обладают своими интересами и предпочтениями, способны самостоятельно выбирать свои действия и пр. [21]?

Одно из проявлений заключается в так называемом манипулировании информацией. В теории коллективного выбора давно известно, что активный субъект сообщает информацию, прогнозируя результаты ее использования, и в общем случае не будет сообщать достоверную информацию [3, 22].

Другой пример — так называемый активный прогноз, когда система меняет свое поведение на основании новых знаний, полученных о себе [23].

Снимаются или усугубляются эти и подобные (примеры — краудсорсинг [24], конформное поведение [25] и др. — см. работу [3]) проблемы в области Big Data?

- Если уж был упомянут принцип дополнительности, то необходимо вспомнить и о «*принципе неопределенности*» в следующем (гносеологическом) варианте [18]: текущий уровень развития науки характеризуется определенными совместными ограничениями на «обоснованность» результатов и их области применимости. Применительно к Big Data принцип неопределенности означает, что существует рациональный баланс между степенью детальности описания состояния интересующей нас системы и обоснованностью тех результатов и выводов, которые мы хотим сделать на основании этого описания.
- Традиционно при построении и эксплуатации информационных систем (будь то корпоративные системы или системы поддержки госуслуг, межведомственного документооборота и т. п.) считается, что содержащаяся в них информация должна быть максимально полной, унифицированной и общедоступной (с учетом разделения прав доступа). Но ведь возможно показывать каждому пользователю реальность, искаженную в своем «кривом зеркале» — создавать для каждого свою индивидуальную информационную картину[20], осуществляя тем самым *информационное управление* [21, 23]. Стремиться к этому или бороться с этим в области Big Data?

ВМЕСТО ЗАКЛЮЧЕНИЯ

Итак, данные всегда были «большими». Интенсивно появляются новые, все более совершенные инструменты их сбора, хранения и обработки. Хотелось бы уметь делать это эффективно и в реальном времени — для этого *необходимо развитие соответствующих отраслей прикладной математики и компьютерных наук* (что является актуальным запросом от технологий и практики к современной науке). Хочется надеяться, что российские ученые (специалисты по ИКТ, прикладной математике, управлению) не окажутся в стороне от этой мировой тенденции — сейчас интенсивно появляются новые журналы [26—28 и др.] и конференции [29—32 и др.], организуемые ведущими мировыми ассоциациями (IEEE и др.), посвященные проблематике Big Data; все больше государственных и коммерческих грантов выделяется по этой тематике.

Также необходима *массовая подготовка специалистов по большим данным, большой аналитике и большой визуализации* (со специализацией в конкретных предметных областях).

Но этого мало — необходимо получение знаний (в рамках соответствующих отраслей науки) и *развитие моделей, позволяющих компактно и адекватно (с учетом решаемой задачи) описывать интересующие нас явления и процессы*. Другими словами, в каждой из областей возможных приложений Big Data *желательно стремиться сделать шаг от «Браге» до соответствующего «Ньютона»*, иначе мы обречены оперировать частностями, за деревьями не видя леса (см. также цитату, вынесенную в эпиграф).

[19] Правда, можно «складировать» данные на всякий случай на будущее — вдруг когда-то возникнет мысль, что с ними делать — захочется, например, проверить ту или иную гипотезу, а данные уже под рукой.

[20] Как минимум — часть «объективной» картины (правду, только правду, но не всю правду), как максимум — произвольную непротиворечивую систему представлений о реальности.

Кроме того, опережающее развитие технологий стало общецивилизационной проблемой, которая должна учитываться как учеными и инженерами, в том числе, в области больших данных, так и потребителями разрабатываемых ими методов и средств.

ЛИТЕРАТУРА

1. *Nature.* — 2008. September 3 (Special Issue).
2. *Dorogovtsev S.* Lectures on Complex Networks. — Oxford: Oxford University Press, 2010.
3. *Губанов Д.А., Коргин Н.А., Новиков Д.А., Райков А.Н.* Сетевая экспертиза. — М.: Эгвес, 2010.
4. *Jackson M.* Social and Economic Networks. — Princeton: Princeton Univ. Press, 2010.
5. *Polyak B.T., Tremba A.A.* Regularization-based Solution of the PageRank Problem for Large Matrices // Automation and Remote Control. — 2012. — Vol. 73, N 11. — P. 1877—1894.
6. *Райгородский А.М.* Модели случайных графов и их применения // Тр. МФТИ. — 2010. — Т. 2, № 4. — С. 130—140.
7. *Сетевые* модели в управлении // Управление большими системами (спец. вып.). — 2010. — Вып. 30.1.
8. *Словохотов Ю.Л.* Физика и социофизика. Ч. 1 — 3 // Проблемы управления. — 2012. — № 1. — С. 2—20; № 2. — С. 2—31; № 3. — С. 2—34.
9. *Губанов Д.А., Новиков Д.А., Чхартишвили А.Г.* Социальные сети: модели информационного влияния, управления и противоборства. — М.: Физматлит, 2010.
10. *Викторов Д.* Большая проблема Big Data в России // Компьютерра. 7 февраля 2013. — URL: http://www.computerra.ru/52840/bolshaya-problema-big-data-v-rossii/ (дата обращения: 8.09.2013).
11. *Algorithmic* Game Theory / Ed. N. Nisan, T. Roughgarden, E. Tardos, and V. Vazirani). — N.-Y.: Cambridge University Press, 2009.
12. *Boyd S., Parikh N., Chu E.,* et al. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers // Foundations and Trends in Machine Learning. — 2011, — N 3 (1). — P. 1—122.
13. *Granichin O.N., Pavlenko D.V.* Randomization of Data Acquisition and l1-Optimization (Recognition with Compression) // Automation and Remote Control. — 2010. — Vol. 71, N 11. — P. 2259—2282.
14. *Nesterov Y.* Efficiency of Coordinate Descent Methods on Huge-scale Optimization Problems // SIAM Journal on Optimization. — 2012. — Vol. 22, N 2. — P. 341—362.
15. *Shoham Y., Leyton-Brown K.* Multiagent systems: Algorithmic, Game-Theoretical and Logical Foundations. — Cambridge: Cambridge University Press, 2009.
16. *Wooldridge M.* An Introduction to MultiAgent Systems. — N.-Y.: John Wiley & Sons, 2002.
17. *Kuhn T.* The Structure of Scientific Revolutions. — Chicago: University of Chicago Press, 1962.
18. *Novikov A., Novikov D.* Research Methodology: From Philosophy of Science to Research Design. — Leiden: CRC Press, 2013.
19. *Simon H.* The Sciences of the Artificial / 3rd Edition. — The MIT Press, 1996.
20. *Garey M., Johnson D.* Computers and Intractability: A Guide to the Theory of NP-Completeness. — San Fancisco. W. H. Freeman and company, 1979. — 14 p.
21. *Novikov D.* Theory of Control in Organizations. — N.-Y.: Nova Science Publishers, 2013.
22. *Aizerman M., Aleskerov F.* Theory of Choice. — Amsterdam: Elsevier, 1995.
23. *Novikov D., Chkhartishvili A.* Reflexion and Control: Mathematical Models. — Leiden: CRC Press, 2014.
24. *Surowiecki J.* The Wisdom of Crowds. — N.-Y.: Anchor, 2005.
25. *Breer V., Novikov D.* Models of Mob Control // Automation and Remote Control. — 2013. — Vol. 74 (in press).
26. *URL:* http://www.journalofbigdata.com (дата обращения: 8.09.2012).
27. *URL:* http://www.hipore.com/ijbd (дата обращения: 8.09.2012).
28. *URL:* http://cci.drexel.edu/bigdata/bigdata2013 (дата обращения: 8.09.2012).
29. *URL:* http://www.ieeebigdata.org/2014 (дата обращения: 8.09.2012).
30. *URL:* http://www.liebertpub.com/overview/big-data/611 (дата обращения: 8.09.2012).
31. *URL:* http://www.swinflow.org/confs/bdds2013 (дата обращения: 8.09.2012).
32. *URL:* https://theinnovationenterprise.com/summits/big-data-innovation-boston (дата обращения: 8.09.2012).

Статья представлена к публикации членом редколлегии Ф.Т. Алескеровым.

Дмитрий Александрович Новиков — чл.-корр. РАН,
зам. директора, Институт проблем управления
им. В.А. Трапезникова РАН, г. Москва,
☎ (495) 334-75-69, ✉ novikov@ipu.ru.