Ю.Н. ТОЛСТОВА

СОЦИОЛОГИЯ И КОМПЬЮТЕРНЫЕ ТЕХНОЛОГИИ

ТОЛСТОВА Юлиана Николаевна – доктор социологических наук, ординарный профессор НИУ ВШЭ (E-mail: untolstova@mail.ru).

Аннотация. *В рамках информационных технологий разработаны компьютерные приёмы решения социологических задач[1], остающиеся неизвестными большинству социологов. Стремления хотя бы приблизительно ознакомиться с достижениями компьютерной науки не наблюдается. Главная причина этого: социологи не знают соответствующих методов. В статье кратко описаны перспективные для социологии технологии, показана их полезность. Описание дано в контексте обсуждения проблемы использования в социологии математических методов, лежащих в корне любой компьютерной технологии.*

Ключевые слова: *информационные технологии • анализ данных • интеллектуальный анализ данных • большие данные • цифровые гуманитарные науки • наука о данных • знание*

Роль компьютера в социологии. Анализ данных. Использовать математические методы социология начала задолго до компьютерной эпохи. Кондорсе и Кетле призывали решать социологические задачи с помощью теории вероятностей [Толстова, 2009б], получали с помощью математических методов интересные социологические результаты. Во второй половине XX в. возможности математического аппарата социологии резко расширились. Родилось понятие *анализ данных (АД)*, которому ниже уделено внимание.

Два вводных замечания. 1. В совокупность методов *АД* обычно включают методы математической статистики, хотя эти группы методов различны. Математическая статистика – строгая ветвь математики (если не считать, что в процессе её применения используется субъективно заданный уровень значимости оценки параметра или проверяемой статистической гипотезы). Собственно *АД* – наполовину математика, наполовину содержательная ветвь науки, поскольку элементы соответствующих алгоритмов выбираются из содержательных соображений, алгоритмы часто эвристичны по характеру и т.д. [Толстова, 2000]. 2. Вместе с методами *АД* развивались методы измерения. Это разные методные ветви, но их часто трудно чётко отделить друг от друга [Толстова, 2009а]. К концу XX в. было разработано много методов измерения и *АД*.

Вычислительная техника доступна практически каждому социологу, плюсы чего очевидны. Главный в том, что развитие IT позволило преодолеть отторжение

[1] Мы не пользуемся термином “эмпирическая социология”, считая порочным делить социологию на эмпирическую и теоретическую. Эти её стороны должны опираться на достижения друг друга. В частности, при проведении эмпирических исследований социолог должен использовать огромное количество предположений, которые иначе, как теоретическими, назвать нельзя [Толстова, 2013].

социологом сложных формул, да и простых при обилии данных. Компьютер рассчитает любую формулу, любой алгоритм данных, которые ему "поставляет" социолог. Использование компьютера стимулировало разработки новых методов; расширен круг применяемых социологами подходов к измерению и анализу данных[2]. Однако компьютер в работе социолога играет и отрицательную роль. Легкость использования компьютерных технологий гасит бдительность при проверке пригодности конкретного метода для конкретной содержательной задачи. О возможном несоответствии заложенной в методе модели изучаемой социологом ситуации, последний, как правило, не думает. «Знай, нажимай кнопки компьютера и получишь "красивые" классификации, латентные переменные, структуры связей и т.д.», – типично рассуждает средне-статистический коллега; отсюда ложные результаты.

Назову проблему сопряжения заложенной в методе модели с содержанием решаемой социологической задачи *первой проблемой АД*. Заметим: одна из часто используемых социологом моделей связана с совокупностью предположений о статистическом[3] характере изучаемого явления. А если такие предположения несостоятельны (что нередко встречается; в частности, не статистическими являются многие методы интеллектуального анализа данных, о которых речь ниже), то использование методов статистического анализа приводит к бессмысленности.

Вторая проблема АД – репрезентативность изучаемой выборки объектов. В социологии нередко невозможно проверить репрезентативность выборки. Расчет её объема по математико-статистическим формулам может быть бессмыслен в силу ряда причин [Толстова, 2007: 90–92]. И эти формулы теряют смысл при отказе от статистичности изучаемой ситуации. Сформулированные проблемы должны грамотно решаться в любом социологическом исследовании, при использовании любых компьютерных технологий.

Человеческая мысль расширяет круг методов (и постановок задач!), глубже изучает закономерности природы и общества. Бурно развивающиеся компьютерная наука и методы измерения и анализа данных (пригодных, в том числе, и для социологии) дают возможность исследователю эффективнее решать задачи в привычных (со второй половины XX в.) постановках, ставить и решать новые классы задач. Опишу четыре компьютерных технологии (подхода), развивающихся в настоящее время и полезных социологу: *Data mining (DM); Большие данные, Big data (BD); цифровые компьютерные науки, Digital Humanity (DH); Наука о данных, Data science (DS)*[4].

Современные компьютерные технологии *(DM, BD, DH, DS)* и расширение возможностей социолога. Строгих определений названных технологий нет. Ни о какой унификации терминологии речи не идёт (какую-то роль могут играть разные области генезиса методов). Примеры синонимов: "искусственный интеллект", "интеллектуальный анализ", "бизнес-аналитика, "бизнес-интеллект". Четкого состава алгоритмов ни для одной технологии не определено (вариантов огромное количество). Рассматриваемые технологии взаимосвязаны, их трудно отделить друг от друга, они переплетены. Но каждой всё же отвечает ядро, отражающее её сущность, фокус соответствующих разработок. Каждое ядро может реализоваться по-разному в разных технологиях, но в рамках отвечающей ему технологии оно реализуется наиболее развернуто.

Иногда совокупность технологий, методов, концентрирующихся вокруг одного ядра, называют парадигмой, что, представляется, имеет право на существование.

---

[2] Нельзя сказать, что традиционные методы измерения и анализа данных в настоящее время используются социологами в масштабах, которые отвечают их полезности для социологии.

[3] Для нас термин "статистический" эквивалент термина "математико-статистический".

[4] Подобных технологий много. Краткое описание их см.: [Давыдов, 2005; 2009]. Указанный автор долго и тщетно пытался привлечь внимание к ним социологов.

Мы будем называть рассматриваемые компьютерные технологии также системами и парадигмами. Опишем ядра рассматриваемых парадигм.

(1) для *DM* – это алгоритмы анализа данных, рассчитанного на более "слабые" и в большей мере соответствующие мышлению человека (относимые к области искусственного интеллекта и отвечающие современному пониманию категории "знание"[5]) модели, требующие меньших предположений относительно анализируемой исследователем ситуации, чем традиционные методы анализа данных; естественно, такие алгоритмы облегчают решение нашей первой проблемы *АД;*

(2) для *BD* – элементы компьютерной науки, позволяющие виртуозно работать с поиском и формированием сложных, больших, разбросанных, меняющихся массивов (баз) данных. Предполагается также, что данные могут быть структурированы и неструктурированы[6]; возможность творческой работы с большими и разнообразными данными значительно облегчает решение второй проблемы *АД;*

(3) для *DH* – это умение превращать неструктурированные данные в структурированные путем оцифровки или моделирования посредством графа (сети);

(4) для *DS* – такая работа с данными, в которой делается упор на понимание последних как носителей скрытых закономерностей, нового знания.

Интеллектуальный анализ данных (Data mining, DM). Подход известен у нас в стране довольно давно, имеются учебники, например [Дюк, 2001; Чубукова, 2006]. Термин введен Пятецким-Шапиро в 1989 г. Иногда его заменяют термином Knowledge Discovery In Databases (*KDD*). Как уже сказано, наряду с традиционными методами анализа данных и математической статистики, рассматриваемая технология содержит алгоритмы, приближенные к логике человека, "вручную" анализирующего данные.

Подобные алгоритмы относят обычно к области искусственного интеллекта. Они требуют относительно мало условий для реализации (методы классификации, основанные на применении ассоциативных правил, деревьев решений, генетических алгоритмов, искусственных нейронных сетей и т.д.). Поэтому как синоним выражения "*Data mining*" часто употребляется выражение "*Интеллектуальный анализ данных*". Хотя традиционные методы анализа данных и математической статистики обычно тоже включают в состав технологии *DM,* надо помнить: заложенные в них модели в большей мере нагружены содержанием и поэтому гораздо менее пригодны для решения задач, на которые рассчитан *DM* (см. ниже).

За счет того, что методы искусственного интеллекта опираются на более слабые модели реальности, при их использовании будет легче осуществляться соблюдение требуемого соответствия формализма содержательным представлениям социолога об изучаемой социологической ситуации. Вероятность того, что формализм будет адекватным для произвольной совокупности данных больше, чем аналогичная вероятность для моделей, предусматриваемых для традиционных методов *АД* и математической статистики. Именно указанное ослабление моделей и дает возможность по-новому ставить задачи. И эта новая постановка обычно считается основной особенностью *DM.* Опишем её и покажем актуальность такой постановки для социолога.

Новая постановка задач предполагает три момента.

1. Исследователь ставит перед собой в качестве цели не решение конкретной задачи (например, построения типологии респондентов, входящих в изучаемую социологом выборку), а выявление того, можно ли в имеющихся у него данных найти какие-то потенциально интересные для социологии закономерности (типологии, структуры

---

[5] Разработке понятия "знание" в литературе уделяется много внимания. Мы не будем определять этот термин.

[6] Определения структурированных и неструктурированных данных мы не даём, полагаясь на интуитивные представления социолога.

связей, яркие визуализации данных). Понятие *закономерности* мы отождествляем с нахождением ранее неизвестного, не очевидного, доступного для интерпретации, полезного *знания.*

2. Подход рассчитан в основном на выявление закономерностей в больших массивах данных (о "больших данных" и их значимости для социологии см. следующий раздел, посвященный описанию *BD*). Предполагается активное использование возможности творческой работы с имеющимися базами данных. Данные комбинируются с целью нахождения массивов, в которых искомое знание действительно содержится.

Первые два пункта вполне отвечают названию *DM:* "добыть руду из неизвестной породы" (mining – горные (рудные) разработки).

3. Ещё одной специфической характеристикой технологии *DM* является то, что, по замыслу её создателей, она должна давать возможность получать новое знание человеку, не разбирающемуся в математике (можно не говорить о важности этого свойства для социолога). Поэтому система содержит эффективные средства визуализации данных и результатов анализа. Однако и на этом примере можно продемонстрировать то, что при использовании, казалось бы, самых простых методов нельзя забывать о необходимости сочетания формализма и содержания (см. первую проблему – *АД*).

Например, в качестве результата визуализации мы можем использовать полигон распределения какого-либо признака. Это – популярное решение задачи описания совокупности объектов; оно требует тщательного выбора признака и способа его операционализации, заполнения пропусков в его значениях, разбиения совокупности этих значений на интервалы, взвешивания значений и т.д. Эти операции можно успешно осуществить, только опираясь на смысл решаемой задачи[7]. Скажем, разбиение диапазона изменения значений признака на интервалы можно осуществить по-разному, результат любого дальнейшего анализа полученного распределения зависит от способа разбиения. Оно может либо опираться на содержательные соображения, либо обусловливаться моделями, которые заложены в известных математических методах разбиения, если мы таковые используем [Толстова, 2000:133–134]. В случае же, когда мы не имеем конкретной задачи и пытаемся понять, есть ли в изучаемой "породе" полезная "руда", то, вероятно, надо опробовать множество разбиений и найти (или не найти) в конце концов то, которое даст результаты. В процессе же дальнейшей работы, конечно, надо учитывать соответствующие интерпретации.

Итак, система *DM* сочетает возможности эффективной работы с базами данных большого объема, с методами искусственного интеллекта, традиционного *АД* и математической статистики. Перейдем к подробному раскрытию первого аспекта.

Большие данные (Big data, BD). Как сказано выше, ядром этой технологии является набор достижений компьютерной науки, позволяющих искать и формировать массивы больших, разбросанных, изменяющихся, структурированных и неструктурированных данных. *Большими данными* будем называть данные, обладающие всеми перечисленными свойствами. Нетрудно поверить в то, что удобная и быстрая работа с этими данными требует специфических компьютерных технологий, которые хотя бы потому обычно не применяются современными социологами, что они, как правило, имеют дело с выборками объемом в 2–3 тысяч (и меньше) респондентов. Освоение технологий работы с *большими данными* требует значительных материальных затрат: стоит ли на это идти? Считаю разумным воспринять этот вопрос как риторический.

[7] Поскольку ниже мы уделяем особое внимание ситуации, когда перед исследователем не стоит конкретная задача, вероятно, следовало бы говорить о такой технологии, когда мы средствами системы получаем много вариантов, выбирая тот, который хорошо интерпретируется.

Ответ на него, конечно, утвердителен. Прежде всего нетрудно видеть, что поиск заранее не очерченных закономерностей требует работы именно с *большими данными.* К необходимости работы с такими данными приводит также желание объединить данные разных исследований, например, при сравнении результатов, для "обогащения" данных одного исследователя данными другого, для сравнения отдельных срезов данных в лонгитюдных исследованиях и т.д. Кроме того, некоторые социологические явления в принципе могут быть выявлены только при большом количестве наблюдений: например, – изучить причины миграционных потоков между разными регионами. Заметим, что использование больших массивов, отражающих разнородные ситуации, позволяет уйти от необходимости строить репрезентативную выборку, от статистических моделей, что в наше время становится актуальным.

Как мы отметили, социологи игнорируют новые IT; трудно в литературе отыскать социологический пример применения *BD*. Но вот пример не социологического характера [Смирнов, 2013]: «Когда в 2009 году началась пандемия гриппа H1N1 (сочетавшего в себе элементы "птичьего" и "свиного"), главная проблема заключалась в том, что органы здравоохранения катастрофически – на одну–две недели – опаздывали с выявлением новых очагов распространения заболевания и, соответственно, с принятием мер.

И тут совершенно неожиданно свое решение предложила компания Google. В общем-то понятно, что определенные поисковые запросы (например, "капли для носа купить"), могут дать представление о количестве заболевших. Но какие именно из трех миллиардов запросов, ежедневно отправляемых пользователями Google, свидетельствуют о самых первых признаках болезни? Специалисты Google разработали специальную систему, которая позволила сравнить графики изменения популярности 50 миллионов наиболее распространенных поисковых запросов с фактическими данными эпидемиологов за 2007–2008 гг. и проверить 450 миллионов математических моделей! В итоге были выявлены 45 поисковых запросов, которые на 97%(!) коррелировали с фактическими данными. Медики получили надежный, работающий в реальном времени инструмент выявления новых очагов распространения гриппа».

Многие гипотетические социологические задачи могли бы быть решены с помощью технологий *BD.* В книге [Pentland, 2014] вводится понятие социальной физики (вспомним Кетле, который совершал этот шаг; наука развивается по спирали) и утверждается, что использование её принципов важно сейчас в силу роста глобальной конкуренции, изменений окружающей среды и недоверия людей к власти. Социальная физика, по мнению автора, отвечает на встающие вопросы с помощью больших данных, возможность работы с которыми делает доступным изучение *всех сторон человеческой жизни.*

Кратко об истории термина *BD. Big Data* относится к числу немногих названий, имеющих вполне достоверную дату своего рождения – 3 сентября 2008 г., когда вышел специальный номер старейшего британского научного журнала Nature, посвященный поиску ответа на вопрос "Как могут повлиять на будущее науки технологии, открывающие возможности работы с большими объемами данных?", в котором собраны материалы о феномене взрывного роста объёмов и многообразия обрабатываемых данных и технологических перспективах в парадигме вероятного скачка "от количества к качеству"; осознавая масштаб грядущих изменений, редактор номера *Nature* К. Линч предложил для новой парадигмы название "*Большие Данные",* выбранное им по аналогии с такими метафорами, как Большая Нефть, Большая Руда и т.п., отражающими не столько количество чего-то, сколько переход количества в качество [Черняк, 2011].

Появление термина *Большие данные* не символизировало нечто абсолютно новое. О необходимости работы с большим объемом информации говорили задолго до

2008 г. Наверное, закономерность тут такова: новый термин не возникает на пустом месте, потребность в нем появляется тогда, когда родилось явление, им описываемое. Однако точного определения термина *Большие данные* до сих пор нет. Он понимается как стихийно обрушившаяся лавина данных, как совокупность новых технологий, радикально меняющих информационную среду, как этап технологической революции. Массовая заинтересованность в соответствующем определении свидетельствует, что, скорее всего, в *больших данных* есть что-то качественно иное, чем то, к чему подталкивает обыденное сознание [Черняк, 2011].

Для прояснения картины несколько примеров. "В качестве определяющих характеристик для больших данных отмечают *"три V"*: объём (*volume*, в смысле величины физического объёма), скорость (*velocity* в смыслах как скорости прироста, так и необходимости высокоскоростной обработки и получения результатов), многообразие (*variety*, в смысле возможности одновременной обработки различных типов структурированных и полуструктурированных данных)" [Канаракус, 2011: 7]. *Большие данные* в рамках IT – это серия подходов, инструментов и методов обработки структурированных и неструктурированных данных огромных объёмов и значительного многообразия для получения воспринимаемых человеком[8] результатов, эффективных в условиях непрерывного прироста, распределения по многочисленным узлам вычислительной сети, сформировавшихся в конце 2000-х годов, альтернативных традиционным системам управления базами данных и решениям класса Business Intelligence.

Чтобы прояснить заключенные в понятии *BD* методологические соображения, сделаю замечание о термине *Business Intelligence (BI)*. Он важен для рассматриваемых нами систем IT, родствен тому**,** что говорилось об искусственном интеллекте. По существу он центральный для социолога, поскольку следуя принципам, заложенным в нём, исследователь может и реализовывать замыслы технического характера, и обеспечить содержательную осмысленность осуществляемой работы, полученных результатов. Сложность перевода указанного термина обусловлена многозначностью английского *intelligence:* способность узнавать и понимать; готовность к пониманию; знания, приобретенные путем обучения, исследования или опыта; действие или состояние в процессе познания; разведка, разведывательные данные. Все эти смыслы должны "работать" при описании *BD.* Единого слова, отвечающего всем им, в русском языке нет. Отсюда – переводы, каждый из которых не вполне удовлетворителен.

Слово "*бизнес*" указывает на происхождение рассматриваемых IT-систем (тот же генезис у многих рассматриваемых понятий). В соответствии с первоначальными определениями термина *BI* (конец 1980-х – 1990-е гг.), это – "процесс анализа информации, выработки интуиции и понимания[9] для улучшенного и неформального принятия решений бизнес-пользователями, а также инструменты для извлечения из данных значимой для бизнеса информации. Надо отметить, что большинство определений

---

[8] Понимание закономерности (обобщения, делающего изыскания наукой) как утверждения, которое хорошо воспринимается человеком, существенно для социологии. Например, компьютерные методы поиска латентных переменных (факторный анализ, многомерное шкалирование и т.п.) предполагают, что количество искомых латентных факторов мало и воспринимается человеком как некое новое знание. Если метод дал 25 факторов, вы вряд ли это примете, будете искать способ сжатия информации, сокращения количества переменных. А 4 фактора – хорошо. Точно границу между "большим" и "малым" указать невозможно.

[9] Подчеркнем: термин "понимание" неслучаен в контексте обсуждения использования математических (компьютерных) методов в бизнесе. Это – понимание, хорошо знакомое социологу и используемое чаще в сочетании "понимающая социология". Такое понимание мы имеем в виду, говоря о соотнесении формализма и содержания. Только Вебер говорил о понимании тех смыслов, которые человек вкладывает в свое поведение, а у нас имеется в виду понимание формализма, точнее, толкование его содержательной роли в получении социологом нового знания.

трактуют “business intelligence” как процесс, технологии, методы и средства извлечения и представления знаний” [Большие данные, 2014]. Мы снова возвращаемся к понятию “*знание”.*

Иногда термин *BI* переводят как “интеллектуальный анализ данных” (мы, следуя традиции, использовали этот термин в описании *DM;* полагая, что рассматриваемые технологии применимы не только в бизнесе, да и родились соответствующие методы в разных дисциплинах). Иногда используются калька *бизнес-интеллект,* термин *бизнес-аналитика.* Способы реализации этих технологий должны каждый раз опираться на научный анализ реальной задачи, стоящий, например, за приведенными выше разъяснениями термина *intelligence,* на анализ, связанный с пониманием данных, методов их изучения, интерпретацией результатов, с получением нового знания. А это дело хлопотное, его не всегда можно сделать в текучке, присущей бизнесу; и термин *BD* родился в научном журнале.

В заключение отметим, что *BD,* как любое научное понятие, развивается, меняется. В данном случае исследователи перестают стремиться к обязательному совмещению в одном исследовании всех “трёх V”. На смену им приходят “нишевые” технологии для узкоспециализированных задач [Смерть больших данных, 2012]. Можно сказать, это разные срезы “больших данных”, только более детальные и направленные. В качестве соответствующих ниш автор указывает, в частности, *Умные данные*, *Науку о данных*. О *Науке о данных* речь ниже. А понятие *Умные данные* кратко рассмотрим.

Строгих определений *умных данных*, насколько нам известно, нет. “Разницу между данными обычными и “умными” можно проиллюстрировать на примере информации о продаже. Так, это может быть длиннющий перечень сведений о недельных объемах продаж (обычные данные), а может быть и график, отражающий пики и спады продаж в рамках определенных временны́х периодов (“умные” данные)” [Что такое “умные” данные, 2014]. Другими словами, здесь *умные данные* – результат описания исходной совокупности объектов с помощью распределения некоторого признака. О специфике такого описания см. выше[10]. В литературе при описании *BD,* говоря о способах анализа данных (мы упоминали, что все рассматриваемые 4 ядра в той или иной мере присутствуют при реализации любой технологии-парадигмы), часто ограничиваются процессом их визуализации.

На Западе технологии *BD* бурно развиваются, щедро финансируются правительствами, которые видят в этих технологиях один из локомотивов развития компьютерных мощностей.

Цифровые компьютерные науки (Digital Humanity, DH). В нулевые годы XXI в. появились термины, обозначающие классы однородных исследований: “гуманитарные вычисления”, “компьютинг в гуманитарных науках”, “гуманитарная информатика”, “цифровые гуманитарные науки”, “электронные гуманитарные науки”, “цифровые исследования в гуманитарных науках”, “кибергуманитарные науки”. Из названий ясно: речь идет об использовании IT. Соответствующее научное направление затронуло социальные науки. Появились термины: “вычислительные социальные науки”, “вычислительная социология”, “электронная социальная наука”, “цифровые социальные науки”, “цифровая социология”, “цифровые социальные исследования”. Мощность указанного потока подтверждается тем, что в мире организованы сотни лабораторий данного профиля; Россия заметно отстает от Запада.

Покажем, что наработки в обрасти *DN* полезны и для социологии. Как нетрудно догадаться по названию, цифровые гуманитарные науки *DH* составляют научную

---

[10] Проблема поиска умных данных выводит на необходимость обсуждения триады (данные, информация, знание). Соответствующие определения для социологии не разработаны. Мы об этом не говорим из-за недостатка места.

ветвь на стыке гуманитарных и компьютерных наук. Она предполагает, что исследователь использует оцифрованные материалы (материалы цифрового происхождения) в рамках методологий гуманитарных наук (история, философия, лингвистика, литература, искусство, археология, культура, музыка и т.д.). Естественно, при этом активно используют компьютерные науки, позволяющие применять компьютерные инструменты (и разрабатывать новые) в сборе данных, информационном поиске, использовать привычные для ряда гуманитарных наук методы анализа данных (в том числе интеллектуального) и математической статистики [Цифровые гуманитарные науки, 2014].

Выделим два таких подхода к получению оцифрованных материалов, которые можно считать измерением, построенном на обобщении принципов теории измерений (об этих принципах можно прочитать в: [Пфанцагль, 1976][11], а об обобщении – в: [Толстова, 2007: 275–283]).

Первый подход предполагает специальную оцифровку интересующих гуманитария документов и используется в тех технологиях из числа перечисленных выше, названия которых содержат термин "гуманитарный". Наиболее часто упоминают использование первого подхода *DH* в одном из канадских университетов [http://around-DH/ shake.ru/ news/3901.htm]; целью было исчерпывающее научное описание карты Лондона эпохи Шекспира. Уникальность проекта в том, что почти все дома и улицы исчезли при пожаре 1666 г. и последующих перестройках. Восстанавливали их по историческим источникам, пьесам, памфлетам, земельным документам, рукописям, реконструкциям. Решение таких задач требует, помимо поиска и оцифровки документов, программного обеспечения, опирающегося на содержательную методологию анализа источников. Оцифровка данных столь объемна, а количество источников велико, что, помимо мощного компьютера, требовалась помощь коллектива аспирантов и студентов. Изыскания привели к возникновению проблемы научного уровня работы при вовлечении в неё десятков пользователей разной подготовки. Представляется, что в социологии подход *DH* может быть эффективным. Оцифровка документов, писем, биографий, результатов неформализованных опросов и т.д. при адекватной обработке может дать новое знание.

Второй подход к получению оцифрованных данных связан с применением *DH* в анализе социальных сетей (включая компьютерные). Здесь наработано много технологий; в их названиях фигурируют термины, однокоренные со словами "социальный" и "социология". В работе: [Вычислительные социальные науки…, 2012] описано исследование 2002 г. ученых компьютерных наук из Карлтонского колледжа в Нортфилде, штат Миннесота. Они искали механизм, помогающий управлять формированием личных отношений: люди пытаются стать друзьями своих друзей, для чего используют социальные компьютерные сети. В рамках этого подхода имеется много наработок, касающихся опросов респондентов онлайн, сбора данных в компьютерных сетях, учета того, что многие общественные движения организуются в Интернет, что на поведение человека влияет наличие у него мобильника и т.д.

В обоих случаях описание каждого подхода сопровождается обсуждением алгоритмов, позволяющих на базе оцифрованного материала получить новое знание (*умные данные*).

Что касается заложенных в технологии *DH* методов анализа данных (напомним: каждая их этих технологий присутствует в любой другой), у них узкий характер – только описание данных, их визуализации.

Работа по оцифровке документов требует международного сотрудничества. В связи с этим упомянем принятый в марте 2011 г. собранием ученых данного науч-

[11] Мы имеем в виду понимание измерения, когда измерительный процесс есть построение математической модели т.н. эмпирической системы, т.е. совокупности интересующих исследователя объектов как носителей определенных свойств.

ного направления “Манифест Digital Humanities” [Дакос, 2011]. Его авторы говорят о необходимости полной открытости работы в области *DH*, грамотного объединения усилий. Думается, это – не пустая фраза. Можно верить, что такие действия обеспечат качественный рост возможностей гуманитарных и социальных наук благодаря расширению области данных, подвергаемых анализу, привлечению многообразных методов, включая естественно-научные (на чем настаивают авторы Манифеста).

В последние годы появляются статьи на русском языке о *DH* и пользе для гуманитариев от применения этих технологий [Журавлёва, 2012а, 2012б, 2013; Можаева, 2013; Погорский, 2014]. Назовем также “Цикл лекций об использовании ИТ в гуманитарных науках”**,** прочитанный профессором Мелиссой Террас (директор Центра цифровых гуманитарных наук Лондонского университета) 13–15 мая 2014 г. в Гуманитарном институте Сибирского федерального университета (Красноярск) [Террас, 2014]. Область работы профессора Террас охватывает и гуманитарные науки, и информационные технологии, и инженерное направление. Она “гибридный специалист”, каких не хватает нашим гуманитарным и социальным наукам. Доклад этот тем более достоин внимания, что, по словам г-жи Террас, она основатель направления *DH* (2004), руководит первой в мире лабораторией, реализующей это направление. По её словам, работа её лаборатории значима воздействием на общество.

Наука о данных (Data science, DS). Что такое *данные?* Вопрос не так прост, как кажется на первый взгляд. Пример данных – набор чисел ответов респондентов на вопросы анкеты. Вдумаемся в ситуацию. Перед исследователем множество чисел; ему известно, что, применяя к этому множеству определенные арифметические операции, есть шанс открыть новые закономерности. Казалось бы, здесь есть нечто мистическое. В чем специфика совокупности чисел? Чем она отличается от других наборов чисел? Как понять, “прячется” или нет в наших данных социальная закономерность? Эти вопросы проходят через описанные выше технологии. Но в *DS* именно на этом сделан акцент. Однако прежде об истории понятия *наука о данных* (*data science*, иногда “*даталогия*” – *datalogy*). Зарождение *DS* относят к 1966 г., к учреждению комитета по данным для науки и техники Международного совета науки (Codata); в России – Национальный комитет КОДАТА по сбору и оценке численных данных для науки и техники; комитет активно работает и сейчас; но обслуживает он естественные науки.

“Наука о данных”, еще без названия, считалась академической наукой [Smith, 2006]. Название появилось в 1990 г., обозначая профессию, которая, как ожидалось, будет извлекать смысл из огромных массивов хранимых данных. С начала 2010-х гг., во многом благодаря популяризации концепции *Больших данных,* новое направление бурно развивается [Dhar, 2013], понимаемое и как научная ветвь, и как практика межотраслевой деятельности. Специализация “учёный по данным” (*data scientist*) с начала 2010-х гг. – одна из привлекательных, высокооплачиваемых, перспективных профессий [Пресс, 2013; Davenport, 2012].

В статье [Пресс, 2013] говорится, что *Наука о данных* – результат “слияния зрелой дисциплины статистического анализа с молодой наукой информатикой”. Главным для неё, как науки, называют извлечение смысла из данных. В связи с этим в названной работе упомянут момент, который нам кажется основным. Речь идет о связи *Науки о данных* с *анализом данных*, с пониманием последнего. *Анализ данных* – это скорее не наука, но искусство, это не математика, но способ получения нового знания, способ, включающий в себя математические формулы моделирования фрагментов реальности, поддающихся моделированию [Толстова, 2000: 95–105]. Примерно то же говорит [Пресс, 2013]. Автор этой статьи вспоминает основоположника анализа данных Д.У. Тьюки, в 1962 г. написавшего в книге “Будущее анализа данных”: “*Долгое время мне казалось, что я специалист в области статистики, заинтересованный в умозаклю-*

чениях, идущих от частного к общему. Но наблюдая за эволюцией математического статистического анализа, я всерьез задумался и начал сомневаться в своем предназначении и призвании... До меня дошло, что в первую очередь мне интересен анализ данных... Анализ данных и те части статистического анализа, которые поддерживают его, должны приобретать черты научного знания, а не математики... анализ данных, по своей внутренней сути – эмпирическая наука".

При использовании *АД* исследователю приходится прибегать к содержательным представлениям об изучаемой ситуации. Необходимость для исследователя проявлять соответствующие способности только растет при переходе к рассматриваемым нами технологиям. Так, автор предисловия к работе классиков анализа данных пишет о необходимости "игры с предпосылками" в процессе анализа данных. Исследователь полагает истинными то одни, то другие предпосылки и смотрит, как это меняет результирующие выводы подобно "добыче руды". А далее говорится, что анализ – это форма существования данных [Адлер, 1982: 7]. Такое положение представляется верным и хорошо отражающим суть анализа данных.

Подобного рода суждения, на наш взгляд, и дают ответ на вопрос о том, что означает выражение "искомые закономерности "прячутся" в данных". Однако этот ответ не полон. Необходимо помнить, что Дж. Тьюки, идеи которого упомянуты выше, – основатель классического *АД*. Как мы отмечали, современные технологии принципиально отличаются от традиционной логики *АД* тем, что рассчитаны на выделение ценной "руды" (знания) из огромного количества разных порций перерабатываемой "породы". Другими словами, в рамках соответствующей парадигмы мы постоянно меняем и методы "добычи" (*анализ данных*), и порции "породы" (анализируемые массивы данных). При этом принципиальны и подбор алгоритмов (они должны в основном опираться на "интеллектуальные" модели), и наличие широкого круга доступных, подходящих для решения конкретных задач баз данных. Последнее заставляет предъявлять специальные требования к организации данных в мировой науке. Соответствующие принципы складываются. Основные среди них два. 1. Исследователи должны участвовать в ширящемся процессе создания баз данных, особенно слабо структурированных и неструктурированных, а также полученных в результате приведения информации из неструктурированного в структурированный вид. Пример такой базы, аккумулирующей извлеченные из 60 библиотек США публикации более 3-х млн записей: книг, статей и т.д., описан в [Погорский, 2014]. Создатели таких баз, кроме компьютерных проблем, решают много организационных вопросов типа преодоления ограничений авторского права. Огромное внимание уделяется способам оцифровки неструктурированных документов. 2. Все используемые базы данных должны быть открытыми *(Open data)*. Примеры соответствующих международных инициатив см. [Там же]. Реализация этих принципов требует международной солидарности ученых. Мы говорили об этом применительно к соответствующей технологии (упомянутый международный манифест); то же касается *DS*. Заметим, что мы не нашли примеров участия ученых России в подобных международных научных движениях. Надеемся, настоящая статья подтолкнет социологов к освоению компьютерных технологий и к требующемуся для успешного их применения международному сотрудничеству.

СПИСОК ЛИТЕРАТУРЫ

Адлер Ю. Наука и искусство анализа данных // *Мостеллер Ф., Тьюки Дж.* Анализ данных и статистика. М.: Финансы и статистика, 1982. С. 5–13.

Вычислительные социальные науки: создание связей. Часть 2. URL: http://mirnt.ru/statji/vychislitelnye-socialnye-nauki-2 (дата обращения: 15.03.2015).

Давыдов А.А. Компьютерные технологии для социологии: обзор зарубежного опыта // Социологические исследования. 2005. № 1. С. 131–138.

Давыдов А.А. Развитие современных интернет-технологий – вызов современной российской социологии. URL: http://www.isras.ru/index.php?page_id=957 (дата обращения: 15.03.2015).
Дакос П.М. Манифест Digital Humanities. URL: http://tcp.hypotheses.org/501 (дата обращения: 15.03.2015).
Дюк В.А., Самойленко В.П. Data mining. Учебный курс. СПб.: Питер, 2001.
Журавлёва Е.Ю. Эпистемический статус цифровых данных в современных научных исследованиях // Вопросы философии. 2012а. № 2. С.113–123.
Журавлева Е.Ю. Развитие исследований в области электронной социальной науки // Социологические исследования. 2012б. № 7. С. 99–107.
Журавлева Е.Ю. К типологии методов интернет-исследования // Вопросы философии. 2013. № 5. С. 84–93.
Канаракус К. Машина Больших Данных // Сети. 2011. № 4. С. 18–26.
Можаева Г.В. Гуманитарные науки в эпоху цифровых технологий: от отраслевой информатики к Digital Humanities // Открытое и дистанционное образование. 2013. № 3 (51). С. 10–16.
Погорский Э.К. Особенности цифровых гуманитарных наук // Информационный гуманитарный портал "Знание. Понимание. Умение". 2014. № 5. URL: http://www.zpu-journal.ru/e-zpu/2014/5/Pogorskiy_Digital-Humanities/ (дата обращения: 27.04.2015).
Пресс Д. Очень короткая история науки о данных // WebScience.ru. 2013. URL: http://webscience.ru/details/ochen-korotkaya-istoriya-nauki-o-dannyh (дата обращения: 15.03.2015).
Пфанцагль И. Теория измерений. М.: Мир, 1976.
Смерть "Больших Данных". Кто на новенького? 2012. URL: http://rtbinsight.ru/articles/the-death-of-big-data-whos-new.html (дата обращения: 15.03.2015).
Смирнов Ю. Большие данные – большие перемены. URL: http://ideas4future.info/2013/11/24/big_data_-_bolshie_peremeny/ 2013 (дата обращения: 15.03.2015).
Толстова Ю.Н. Анализ социологических данных. М.: Научный мир, 2000.
Толстова Ю.Н. Математико-статистические модели в социологии. М.: ИД ГУ-ВШЭ, 2007.
Толстова Ю.Н. Измерение в социологии. М.: ИД ГУ-ВШЭ, 2009а.
Толстова Ю.Н. Сущность математики в преломлении к потребностям социологии: уроки истории // Математическое моделирование социальных процессов. Вып. 10. М.: КДУ, 2009б. С. 376–423.
Толстова Ю.Н. История методов социологического исследования как отражение эволюции теоретической мысли в социологии // Социологические исследования. 2013. № 8. С.13–23.
Террас М. Цикл лекций об использовании *ИТ* в гуманитарных науках.
Лекция 1. «Что такое "цифровые гуманитарные науки"?» (13 мая 2014 г.) URL: http://tube.sfu-kras.ru/video/1793?playlist=1792 (дата обращения: 15.03.2015).
Лекция 2. "Цифровые гуманитарные науки и участие публики: как включить общую публику в изучение культуры с помощью цифровых инструментов". *М. Террас* (14 мая 2014 г.). URL: http://tube.sfu-kras.ru/video/1794 (дата обращения: 15.03.2015).
Лекция 3. Новые подходы к оцифровке: как создавать изображение (15 мая 2014 г.) URL: http://tube.sfu-kras.ru/video/1795 (дата обращения: 15.03.2015).
Что такое "умные" данные? 2014. URL: http://eagi.kz/index.php?dn=article&id=2324&to=art (дата обращения: 15.3.2015).
Черняк Л. Большие Данные – новая теория и практика // Открытые системы. 2011. №10. URL: http://www.osp.ru/os/2011/10/13010990 (дата обращения: 15.03.2015).
Чубукова И.А. Data mining. Учебн. пос. М.: Бином, 2006.
Davenport T.H., Patil D.J. Data Scientist: The Sexiest Job of the 21st Century // Harvard Business Review. 2012. № 1. URL: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/ (дата обращения 27.04.2015).
Dhar V. Data Science and Prediction // Communications of the ACM. 2013. Т. 56. № 12. P. 64–73.
Pentland A. Social Physics: How Good Ideas Spread: The Lessons From a New Science. New York: The Penguin Press, 2014.
Smith F.J. Data Science as Academic Discipline // Data Science Journal. 2006. V. 5. URL: http://around-DH/ shake.ru/news/3901.htm (карта Лондона) (дата обращения: 15.3.2015).