Большие данные в социологии: новые данные, новая социология?*

Катерина Губа
Кандидат социологических наук, младший научный сотрудник
Института проблем правоприменения
Европейского университета в Санкт-Петербурге
Адрес:. ул. Шпалерная, д. 1, г. Санкт-Петербург, Российская Федерация 191187
E-mail: kguba@eu.spb.ru

В статье рассматриваются перспективы использования больших данных в социологии. В социальных науках сравнительно недавно появились призывы анализировать поведение человека при помощи новых способов производства, обработки и методов анализа данных. Особое внимание к новым данным характерно в первую очередь для социологии, для которой они могут означать переориентацию всего проекта дисциплины. Дискуссия о концептуальных особенностях больших данных развивалась от обсуждения их размера к пониманию, что их ключевая черта — в способе производства. Отличие новых данных состоит в том, что они создаются не для целей исследования, охватывают всю популяцию и производятся в режиме реального времени. Масштаб данных о поведении людей на микроуровне меняет те научные области в социологии, для которых ранее существовали серьезные ограничения при исследовании социального поведения. Это главным образом позволило продвинуться в решении теоретической проблемы, связанной с определением природы социального влияния. В свою очередь, заимствования инструментов из компьютерной науки изменили способ анализа больших неструктурированных массивов текста, что особенно важно для научных областей, исследующих символическое производство. В социологии культуры новые методы анализа текстовых данных позволяют преодолеть асимметрию — развитие теории всегда опережало развитие методов. Социология меняется с приходом новых данных не только в своих отдельных областях, но и в общем видении дисциплинарного проекта. Статья заканчивается обсуждением новой эмпирико-ориентированной социологии, которая идет от данных и этим отличается от привычного для мейнстрима стиля с последовательным движением от гипотез к сбору и анализу данных. Исследования и раньше далеко не всегда запускались теоретически обоснованным вопросом, однако именно сейчас отсутствие сцепки «теория—данные» формулируется как легитимный вариант дисциплинарного будущего социологии.

Ключевые слова: большие данные, вычислительная социальная наука, сетевой анализ, тематическое моделирование, доказательная социальная наука

Историю социальных наук можно рассматривать как смену этапов, связанных с характером доминирующих данных (Mohr et al., 2013: 676). Весь прошлый век нашими источниками данных служили опросы, интервью и наблюдения. Сейчас начался следующий этап, в котором решающую роль играют новые технологии про-

© Губа К. С., 2018
© Центр фундаментальной социологии, 2018
DOI: 10.17323/1728-192X-2018-1-213-236
* Исследование выполнено за счет гранта Российского научного фонда (проект № 17-18-01618).

изводства и сбора больших данных о тех аспектах поведения человека, которые раньше не поддавались наблюдению. Большие данные появляются не в результате проведения опросов или интервью, их создание опосредовано технологиями: мобильные телефоны, электронная почта, сервисы онлайн-банка, транзакции по кредитным картам, перемещение по сайтам, считывание бар-кодов, социальные сети и т. д. Новый лейбл «вычислительная социальная наука» (computational social science) все чаще используется для обозначения исследовательского поля, в рамках которого поведение человека анализируется при помощи новых способов производства, обработки и методов анализа данных (Lazer et al., 2009). Вокруг новых идей развивается инфраструктура: запущено специальное финансирование, открываются новые магистерские программы, создаются журналы и исследовательские центры.

В настоящем обзоре предпринята попытка ответить на вопрос о том, какие изменения привнесли новые данные в социологию. Если пойти по простому пути, то можно рассмотреть, как новые возможности уже используются в социологии. Другой путь состоит в том, чтобы обратиться к изменениям, затрагивающим весь дисциплинарный проект социологии. В этом случае мы не говорим о тех исследовательских областях, для которых большие данные открыли второе дыхание, но попытаемся предугадать, изменится ли сама дисциплина. Социологи предлагают заменить традиционную социологию на доказательную социальную науку, которая отличается от привычного для мейнстрима стиля с последовательным движением от гипотез к сбору и анализу данных. В статье раскрыты оба способа ответа на вопрос о ключевых изменениях в социологии с приходом больших данных.

Вначале мы обратимся к дискуссии о том, что составляет концептуальные особенности больших данных, которые позволяют называть их не столько большими, сколько новыми данными. В следующей части статьи речь пойдет о тех областях социологии, в которых уже отмечен заметный интерес к новым данным. Так, большие данные оказались важными для социологии в двух отношениях. Во-первых, они предоставили возможность изучать социальное поведение, доступ к которому раньше был ограничен (McFarland, Lewis, Goldberg, 2015). Работа с онлайн-данными позволила продвинуться в решении теоретической проблемы, связанной с определением природы социального влияния — передачи через социальные связи паттернов поведения, установок, болезней или даже эмоций. Во-вторых, заимствование инструментов компьютерной науки изменило способ анализа больших неструктурированных массивов текста, что особенно важно для тех научных областей, где исследуется символическое производство. В заключительной части статьи речь пойдет об изменениях в исследовательском стиле социологии.

Наш обзор имеет свои ограничения. Он посвящен возможностям применения больших данных именно в социологии и не затрагивает другие дисциплины[1]. Мы

1. В экономике смотрите обзор (Einav, Levin, 2013), в менеджменте и бизнесе (Frizzo-Barker et al., 2016), урбанистике (Koonin, Holland, 2014), антропологии (Сивков, 2017), истории (Bearman, 2015).

также не касаемся ограничений, которые связаны с применением больших данных, в том числе этического характера[2].

О природе больших данных

Социальные науки далеко не сразу поддались очарованию новых возможностей — первые научные статьи появляются только в 2009 году. К этому времени эра больших данных уже была провозглашена массовыми изданиями, что связано с появлением нового рода деятельности — аналитики данных в коммерческом секторе (Manovich, 2011). Количество статей в массовой периодике до сих пор заметно превышает количество статей в научных журналах. Большие данные оказались полезными в коммерческом секторе для улучшения рабочих операций и извлечения большей прибыли, поскольку предоставили возможность предсказывать поведение людей на основе уже существующих данных (Goel et al., 2010). Анализируя то, что интересует людей в данную минуту — на каких сайтах они проводят время и какие запросы посылают в поисковые системы, можно предсказать, чего они захотят в ближайшем будущем.

Современную историю больших данных иногда начинают со слов исследователей NASA, которые в 1997 году столкнулись с тем, что их компьютеры не справляются с объемом данных. Таким образом, изначально акцент делался на объеме: большие данные — это данные, с которыми не справляется Excel на простом компьютере. Однако если считать объем за главный параметр, то придется признать относительный характер больших данных, так как возможности обработки больших массивов информации постоянно совершенствуются (Austin, Fred, 2016). В дальнейшем речь стала идти о трех характеристиках — размере, скорости накопления и разнообразии (volume, velocity, variety). Согласно Р. Китчину, большие данные отличаются большим объемом; высокой скоростью накопления (они создаются здесь и сейчас и их объем может увеличиваться каждую секунду); многообразием формы; исчерпывающим характером (зачастую представляют всю совокупность); высокой дискретностью (что позволяет дробить данные на отдельные группы и легко их идентифицировать); возможностью привязки к другим типам данных; гибкостью (добавлять новую информацию и расширять объем) (Kitchin, 2014: 2).

Несмотря на некоторые попытки описать ключевые характеристики больших данных, мы можем говорить скорее о лейбле, который схватывает самые разные данные в одном наименовании (Kitchin, McArdle, 2016). Этот термин потерял концептуальную ясность, что хорошо показано на примере анализа Р. Китчином и Дж. Макардлом 26 наборов данных, использованных в научных исследованиях. Данные под одним лейблом имеют как общие, так и отличные друг от друга характеристики. Более того, исследователи не обнаружили ни одного набора данных,

2. Основные критические аргументы можно найти в: Boyd, Crawford, 2012; Zwitter, 2014; Iliadis, Russo, 2016.

который описывался бы через все семь ключевых характеристик. Они определили только две ключевые черты, которым соответствовали все 26 исследований, — это скорость накопления и всеобъемлющий охват (вся реальность объектов этого типа, n = all). Под эти характеристики подпадают главным образом онлайн-данные или данные, которые создаются за счет электронных технологий. Все данные, которые анализировали Китчин и Макардл, так или иначе, предполагали использование электронных средств. Есть только одно исключение из этого списка — административные данные, которые генерируются государственными ведомствами. При всем отличии от онлайн-данных их, тем не менее, можно считать большими данными в силу того, что они обычно охватывают всю популяцию и производятся в реальном времени, хотя при этом доступ к ним может предполагать временной лаг (Connelly et al., 2016).

Итак, мы видим, что размер не является необходимым условием для определения сути больших данных. Действительно, и прежде существовали данные, которые были достаточно большими, например, данные переписи, с одной стороны, содержали информацию о тысячах единиц, но с другой — не были гибкими, их трудно было дополнить другими данными, и требовались специальные усилия по их генерации (Kitchin, 2014). Исследователи полагают, что революция в больших данных произошла не потому, что теперь можно иметь дело с данными большого объема, главное, что данные создаются не для целей исследования (Connelly et al., 2016: 2). Раньше они собирались по запросу исследователя, зачастую по заранее определенной процедуре и в соответствии с исследовательскими допущениями или гипотезами. Сейчас данные производятся самими пользователями: люди пишут посты, ставят лайки, загружают фотографии и делают покупки, в свою очередь, государственные ведомства становятся обладателями данных о самых разных областях — образовании, медицине, криминологии (Волков, Скугаревский, Титаев, 2016). Сбылась мечта социолога — добраться до следов, которые остаются от действий людей, независимо от намерений исследователей. Причем «больше нет необходимости выбирать между количеством единиц в наших данных и количеством информации о них… Детальная информация и понимание, которое раньше можно было получить только о немногих, сейчас доступны для большого количества людей» (Manovich, 2011).

Итак, создание новых данных почти никак не связано с намерениями ученых провести исследование. Они сочетают два аспекта, которые раньше практически не встречались вместе, — это масштабные данные о поведении людей на микроуровне. Что же нового это означает для социологии?

Новые данные: социальное влияние

Манифестом «новой науки» можно считать статью «Вычислительная социальная наука», которая появилась в «Science» в 2009 году (Lazer et al., 2009). Авторы статьи не раз выступали в роли ключевых спикеров крупных профильных конфе-

ренций, они руководят центрами и институтами, результаты их исследований появляются в престижных «Science» и «Nature»[3]. Исходный тезис: большие данные не могут не изменить социальную науку, поскольку данных такого масштаба на уровне тонких взаимодействий раньше не было. Идея вызвала интерес со стороны представителей разных дисциплин. Информация о ссылках на статью позволяет сделать вывод о том, что внимание к ней обеспечивают ученые, публикующиеся в одних из самых престижных журналов: «Plos One», «PNAS», «Scientific Reports», «Science» (вместе — треть всех статей). При этом особый интерес можно отметить именно у социологов — на это указывает список десяти самых цитируемых журналов в статьях, которые ссылались на «Вычислительную социальную науку». Этот список состоит во многом из междисциплинарных журналов, однако среди дисциплинарных на первых местах находятся социологические издания[4].

В первую очередь социологи видят преимущества в возможности продвинуться за счет новых данных и способов их анализа. Гэри Кинг считает, что наибольший результат в социальной науке возможен, когда присутствуют три условия: инновационные статистические методы, новая компьютерная наука и оригинальные теории отдельных областей знания (King, 2013). В этом смысле социальные науки должны заниматься всем тем же самым, но с лучшими методами и лучшими данными, которые позволяют преодолеть недостатки прежних данных — их искусственные условия создания, ретроспективный характер и статичность собираемой информации (Golder, Macy, 2014).

> Вместо того чтобы пытаться каждые два года извлечь мнения о политике у нескольких тысяч активистов путем искусственно созданной ситуации разговора в виде опросного интервью, мы можем использовать новые методы и получить десятки миллионов политических мнений, которые появляются ежедневно в блогах. Также как вместо того, чтобы изучать влияние контекста на взаимодействия людей, спрашивая респондентов об их последних контактах, мы можем собрать информацию за длительный промежуток времени об их телефонных звонках, письмах и сообщениях. При отсутствии официальной статистики мы можем судить об экономическом развитии и росте населения, основываясь на информации со снимков спутника об освещенности, расположении дорог и других объектов инфраструктуры. (King, 2009: 92)[5]

3. Например, А. Пентланд возглавляет в MIT лабораторию по изучению динамики поведения человека. А.-Л. Барабаши, известный исследователь сетей и создатель отдельного направления науки о сетях, возглавляет Центр исследований сложных сетей. Широкой публике должен быть известен Н. Кристакис, автор бестселлера «Связанные одной сетью. Как на нас влияют люди, которых мы никогда не видели» (в соавторстве с Дж. Фоулером). В Йеле Кристакис возглавляет Лабораторию по исследованиям природы человека, а также Институт исследований сетей. Нужно отметить, что в этой звездной компании мы видим микс социальных ученых и тех, кто не получал степени в социологии или политической науке. Из 15 авторов статьи половина имеет степени по социальным наукам, остальные писали диссертации в области физики и компьютерных наук.

4. На основе данных Web of Science — «American Journal of Sociology» (287), «Social Networks» (268), «Annual Review of Sociology» (142), «American Sociological Review» (130), «Organization Science» (125).

5. Статья Г. Кинга и М. Робертса о цензуре в Китае является хорошим примером, как без такого рода данных было бы сложно провести исследование. В предшествующих исследованиях использова-

Новые данные тем самым расширяют пространство и дают новые возможности для развития привычных направлений исследований, в особенности тех, которые смогут воспользоваться онлайн-данными. Сбор данных о поведении людей в самых разных контекстах сложен, требует ресурсов, а в ряде случаев проблема доступа так серьезна, что некоторые исследовательские вопросы остаются незаданными. Онлайн-данные предоставляют информацию о поведении людей в реальном времени, фиксируя автоматически, кто, где и с кем сейчас взаимодействует; при этом минимизируется влияние исследователя при самом производстве данных, ведь они существуют независимо от того, будет ли он их анализировать или нет (Golder, Macy, 2014). Есть мнение, что онлайн-данные изменили социальные науки, так же как в свое время электронный микроскоп или МРТ изменили естественные науки — новые инструменты позволили наблюдать за онлайн-активностью, и именно это производит трансформирующий эффект на социальную науку (Golder, Macy, 2014).

Главный результат для теоретической социологии, по всей видимости, состоит в появлении возможности использовать онлайн-данные для изучения социального «заражения» — передачи через социальные связи паттернов поведения, установок, болезней или даже эмоций (Christakis, Fowler, 2013). Использование эпидемиологической метафоры требовало уточнения, какой механизм задействован при «заражении». Действительно какие-то вещи — микробы или информация — могут передаваться от человека к человеку без особого участия сетей, а просто через мимолетные контакты. Однако для распространения поведения может быть необходимо у сети наличие определенных структурных характеристик (Smith, Christakis, 2008: 412). Возникла дискуссия о том, какой тип связей необходим для распространения феноменов по сетям. Классические работы показали важность слабых связей или структурных дыр в сетях, которые нужны для перемещения и материальных ресурсов, и информации (Granovetter, 1973; Burt, 2004). Слабые связи имеют ту особенность, что далеко простираются, а значит, могут достичь большего числа людей, в отличие от сильных связей, которые имеют тенденцию к кластеризации. Однако остается вопрос: достаточно ли контакта через слабые связи, чтобы произошла успешная передача? Другая гипотеза заключалась в том, что социальное поведение подразумевает сложное влияние: людям обычно нужно вступить в контакт с несколькими источниками «инфекции», прежде чем они по-

лись государственная статистика, опросы населения и интервью с представителями власти. Каждый из этих источников имел свои серьезные недостатки, чтобы можно было полагаться на них при изучении цензуры. Исследование Кинга строилось на анализе цензуры онлайн-записей. Сбор данных предполагал извлечение с многочисленных сайтов записей, пока они не были прочитаны цензорами и удалены (всего было собрано 3 674 698 записей). В дальнейшем отслеживалось, было ли произведено вмешательство цензора (это случилось в 13 %). Категории, которые цензурировались, — события, относящиеся к коллективному действию, критике цензоров и порнографии, тогда как категории, в которых обсуждались решения правительства, не проходили через цензуру. Государственные лидеры вряд ли рады критическим замечаниям, однако это не тревожит их до такой степени, чтобы задействовать цензуру для удаления критических записей. Что их действительно тревожит, так это события, которые могут способствовать сплочению людей (King, Roberts, 2013).

чувствуют себя готовыми перенять поведенческие образцы (Centola, 2011; Centola, 2010). Тогда успех в социальном «заражении» обеспечивается скорее сильными связями, которые «избыточны» по своему характеру.

Возможность анализировать большие сети позволила продвинуться в этом направлении. Искать ответы прежними способами было слишком дорого. Опросные инструменты строятся на выборочных механизмах, тем самым изначально связи между людьми ограничиваются дизайном исследования — уже нельзя задаться вопросом, как организовано влияние в широких сетях (Golder, Macy, 2014). Социальные ученые в сетевом анализе чаще использовали этнографические методы, собирали информацию для анализа статических сетей небольшого масштаба (McFarland, Lewis, Goldberg, 2015). Если и получалось собрать данные о связях, чтобы измерить социальное влияние, то издержки не позволяли включить в анализ значительное количество случаев. Элизабет Ботт (Bott, 1955) проводила вечера за длинными беседами, чтобы зафиксировать интенсивность связей и взаимодействий, когда изучала лондонские семьи (всего в исследовании участвовали 18 семей). В свою очередь, Б. Уэллману (Wellman, 1979) удалось собрать гораздо больше сетевых данных о канадских рабочих, однако опрос давал возможность задать довольно короткий список вопросов. В учебниках по сетевому анализу присутствуют рекомендации, что респондентов стоит просить предоставить информацию об их связях не более чем с пятью людьми.

Исследователи вынуждены были строить гипотезы, которые учитывали социальное влияние через сильные связи отдельных людей, что не позволяло схватывать структурные характеристики сетей (информацию о том, как выглядит вся сеть контактов человека). Сейчас же в силу того, что взаимодействия оставляют свой онлайн-след, есть возможность собрать для большого количества людей информацию об их телефонных звонках, электронных письмах, смс-сообщениях, адресных книгах и онлайн-взаимодействиях в социальных сетях (King, 2009). Другими словами, появились инструменты, которые настолько облегчили сбор данных об отношениях, что в своем роде открытие науки о сетях произошло заново.

Появление социальных сетей и регистрации онлайн-поведения сделали возможным проводить онлайн-эксперименты, которые позволяют контролировать большую часть условий, а также оценить эффект вмешательства на больших выборках индивидов (McFarland, Lewis, Goldberg, 2015). Один из самых выдающихся примеров основан на данных более 60 миллионов пользователей Фейсбука (Bond et al., 2012). Эксперимент тестировал возможности социального влияния при политической мобилизации на примере выборов в Конгресс. В эксперименте 2010 года участвовали все американские пользователи Фейсбука с 18 лет, которые были разделены на три группы. Первой группе (60 055 176) в новостной ленте показали «социальное сообщение»: в нем содержался призыв проголосовать, информация о месте голосования, а также показывались профили людей из числа друзей пользователя, которые уже проголосовали. В ленте второй группы (611 044) появилось информационное сообщение, в котором также была кнопка, на которую можно

было нажать и тем самым показать друзьям, что ты проголосовал; здесь также присутствовала ссылка о месте голосования, однако социальная составляющая сообщения отсутствовала. Третья группа — контрольная — вообще не получила никаких сообщений. Действия, которые потом анализировались: нажатие ярлыка «I vote», переход по информационной ссылке и голосование на выборах.

Согласно результатам, получившие социальное сообщение чаще нажимали на кнопку о том, что они проголосовали, чем те, кто получил только информационное сообщение. Этот результат повторился и на данных о реальном голосовании, то есть люди из первой группы чаще голосовали, чем пользователи из других групп эксперимента. Между контрольной группой и группой только с информационным сообщением вообще не было статистически важных различий. Это говорит о том, что одна только информация не особенно меняет поведение людей, тогда как социальное давление оказывается действенным, меняя поведение о политическом самовыражении (рассказать друзьям о том, что я проголосовал) и участии в реальном голосовании. В эксперименте также попробовали проверить, насколько влияние зависит от силы связи (частота взаимодействий в виде обмена сообщениями). Оказалось, что близкие друзья человека, который сам нажал на кнопку, также чаще сами нажимали на кнопку «I vote» и чаще голосовали, чем близкие друзья тех, кто был в контрольной группе. Остальные друзья, которые формировали с пользователем слабые связи, оказались не затронутыми влиянием — они не стали чаще сообщать о том, что проголосовали, так же как они не стали чаще действительно голосовать. Результаты подтверждаются и в других онлайн-экспериментах (Covielo et al., 2014).

В обзоре «Annual Review of Sociology» подводятся итоги о вкладе данных об онлайн-поведении в развитие социальных наук (Golder, Macy, 2014). Огромный всплеск интереса зачастую оборачивается исследованиями низкого качества. Многие статьи повторяют идеи предшествующих авторов, но на более масштабных данных, при этом без всякой отсылки к классическим работам. Создается впечатление, что сети дают такой инструмент, который позволяет построить графы почти на любых данных, что лишает его осмысленности (boyd, Crawford, 2012). Вместе с тем именно анализ больших сетей встраивается в классические социологические сюжеты о природе социального влияния, которые при этом исполняются на высоком методологическом уровне. Это направление исследований соединяет все три составляющих, о которых писал Кинг, — инновационные статистические методы, новая компьютерная наука и оригинальные теории отдельных областей знания.

Далее мы увидим, что новые возможности для социологии появляются не только с доступом к данным, которые раньше были слишком дорогостоящи или вовсе отсутствовали, но и с развитием инструментов работы с данными, доступ к которым существовал всегда.

Новые методы анализа текстовых данных

Новые данные превосходят старые в своем объеме, разнообразии и глубине, но обычно они существуют совсем не в том виде, в котором готовы для анализа. Превращение сырых данных в нужный для исследователей формат требует специальных компетенций из области компьютерной науки. Исследователи перечисляют целый арсенал методов: математическое и статистическое моделирование; динамический анализ сетей; автоматическое генерирование гипотез; методы интеграции мультимодальных данных; возможности обработки естественного языка и машинное обучение (Golder, Macy, 2014). Социологам нужны люди, которые умеют программировать не только для того, чтобы извлекать данные, но и для того, чтобы их анализировать. Решение этих задач привело к успехам — авторы пишут о четырех прорывах в анализе данных. Первый связан с большими массивами текстовых данных и отсылает к области вычислительной лингвистики, второй развивает сетевой анализ, третий опирается на достижения машинного обучения, наконец, четвертый использует возможности онлайн-экспериментов (McFarland, Lewis, Goldberg, 2015). Из этого списка особенного внимания заслуживают инструменты из области вычислительной лингвистики. Появление тематического моделирования описывается как шаг революционного значения, который на данный момент пока не оценен социологами в должной мере (Evans, Aceves, 2016). Главные области применения — это социология науки и социология культуры, ведь именно в этих областях исследователи имеют дело с текстами.

Первая социологическая работа, в которой использовалось тематическое моделирование, относилась к социологии науки (Moody, Light, 2006). Однако сейчас мы видим, что основной интерес к этому методу присутствует в социологии культуры. Исследователи считают, что социология культуры всегда отличалась тем, что развитие теории опережало развитие методов:

> Социологи, которые изучают культуру, сформулировали многочисленные теоретические гипотезы и концепты, которые обещают глубокое понимание культурных изменений, но им все еще не хватает инструментов для операционализации концептов. Мы предполагаем, что с помощью тематического моделирования будет возможно операционализировать такие ключевые концепты, как фреймирование, полисемия, гетероглоссия и реляционный характер значений. (DiMaggio, Nag, Blei, 2013: 571)

Вероятно, в силу того, что смыслы всегда методически изучать гораздо сложнее, в социологии культуры был период, когда исследовалась не столько сама культура, сколько то, как она производится (Peterson, Anand, 2004). Отсылающие к смыслам концепты — символические границы (М. Ламонт), культурные инструменты (Э. Суидлер), когнитивные схемы (П. ДиМаджио) и культурные фреймы (Р. Бенфорд и Д. Сноу) — развивались на основе «маленьких» данных, которые

подразумевали технику «медленного чтения» (close reading) транскриптов интервью и проведение контент-анализа ключевых текстов (Bail, 2014).

Действительно, обычно социологи анализировали тексты тремя способами (DiMaggio, Nag, Blei, 2013: 577). Первый основан на интерпретативном чтении, без какой-либо формализации. Второй способ строится на контент-анализе, при котором исследователь заранее создает систему категорий и кодов, согласно которым затем кодируется текст. Ограничением метода оказывается трудоемкость, что делает его малопригодным для анализа большого корпуса текста. При этом заранее нужно хорошо представлять, что можно найти в тексте (DiMaggio, Nag, Blei, 2013: 577). И, наконец, третья стратегия заключается в том, чтобы с помощью программы определить набор ключевых слов, а затем сравнить, как часто в разных частях текста встречаются эти слова. Эта стратегия не совсем устраивала именно социологов, которые изучают культуру, так как слова извлекались без учета смыслового контекста, в который они встроены. Две последние стратегии в большей степени подходят для анализа небольших корпусов текстов, с заранее продуманными вопросами (DiMaggio, Nag, Blei, 2013: 577). Нужен был новый метод, лишенный недостатков прежних. Таким методом, по мнению социологов, является тематическое моделирование, так как именно оно отвечает условиям анализа больших массивов текста.

В чем его преимущества? Этот подход носит эксплицитный характер, то есть массив данных доступен для всех, и анализ можно воспроизвести; подход является автоматическим, что дает возможность работать с текстами больших объемов; он позволяет обрабатывать текст до заранее разработанной схемы; принимает во внимание реляционный характер значений. В рамках тематического моделирования корпус текста автоматически кодируется по нескольким категориям, которые называют темами (topics). Алгоритм может это делать при минимальном участии человека, тем самым метод является индуктивным по своей природе: «Вместо того чтобы начать с заранее определенных смысловых кодов или категорий (как те, которые мы создаем, когда вручную кодируем текст), исследователь задает количество тем, которые должен найти алгоритм. Программа затем находит это заданное количество тем и показывает вероятности слов, используемых в теме, так же как предоставляет распределение тематик по всему корпусу текста» (Mohr, Bogdanov, 2013: 546).

При этом не требуется предварительное близкое знакомство с текстом или заранее разработанная схема кодирования. Инструмент сам создает кластеры, скрытые темы на основе статистических моделей. Сохраняется контекстуальность, так как слова приписываются кластеру на основе их появления рядом с другими словами, так же как и многозначность смыслов, так как слова могут одновременно принадлежать разным кластерам (Mutzel, 2015: 2). Несмотря на то что для такого исследования требуются знания в компьютерной науке и статистике, их невозможно проводить без человека, знакомого с той областью, к которой относится текст. Тем самым в использовании новых инструментов для анализа текста наблю-

дается фундаментальное смещение с предварительной работы по созданию категорий и системы кодирования к интерпретации постфактум, которая запускается, когда алгоритм нашел тематические категории и нужно решить, имеют ли они какое-либо значение. В этом и заключается важное преимущество такого метода, ведь если сначала происходит разработка системы кодирования, то когда она закончена и начался сам анализ текста, сложно вернуться обратно (Mohr, Bogdanov, 2013: 562). Тем самым исследователь значительно более ограничен в процедуре, и ему обязательно нужно глубокое знание поля еще до того, как начать анализ. Кроме того, с новыми инструментами исследователь может найти тематические категории, про которые он и не думал, что они присутствуют в тексте, — в контент-анализе такой возможности нет. Соответственно, есть возможность исследовать и открывать новые паттерны (DiMaggio, Nag, Blei, 2013).

Конкретные области в социологии культуры, которые могут получить развитие в связи с появлением больших данных и новых техник анализа, перечислены в статье К. Бейла. Среди них — картографирование культурного окружения или систем значений, классификация культурных элементов (таких как фреймы или схемы внутри систем), прослеживание изменений в культурных процессах за длительный период времени. Многие вопросы в рамках социологии культуры требуют макроанализа, то есть взгляда сверху на все культурное пространство. Бейл призывает активно пользоваться онлайновыми данными в рамках социологии культуры, так как зачастую в руках исследователя могут оказаться не только текстовые данные, которые интересуют как совокупность каких-то значений, но и социальная информация об акторах, что позволяет ставить более интересные вопросы (Bail, 2014).

Важно понимать, что использование тематического моделирования — это зачастую только начало. Как пишут социологи: «В анализе культуры целью моделирования является понимание структуры данных, чтобы иметь возможность выявить тематические кластеры («голоса» или «фреймы»), которые основываются на данных и поддаются интерпретации. В дальнейшем ученые могут использовать их для постановки более фокусированных вопросов» (DiMaggio, Nag, Blei, 2013: 602–603). Например, в процитированной работе П. ДиМаджио и его коллег анализировалось, как в газетных статьях представлено искусство. Тематическое моделирование позволило увидеть темы, затем исследователи задались вопросом о связи фреймирования искусства в массовой прессе с разнообразными способами его финансирования. Для ответа на этот вопрос уже понадобились техники регрессионного анализа.

Инструменты компьютерной науки развивают новые методы анализа, которые вовсе не обязательно применять только на больших данных. Они могут дать интересные результаты даже на сравнительно маленьких данных, которые раньше анализировались традиционными методами. Например, исследователи считают, что компьютерный анализ лучше работает, чем интерпретативное чтение. Статья Дж. Мора и его соавторов иллюстрирует применение возможностей компьютер-

ного анализа текста к данным небольшого масштаба (Mohr et al., 2013). Они предлагают обратиться к новой стратегии компьютерного чтения текстовых сообщений с использованием аналитической модели, разработанной на основе концептов Кеннета Берка. В исследовании анализируются тексты о стратегии национальной безопасности США с 1990 по 2010 год — это открытые документы, которые публикуются ежегодно. Авторы искали в тексте структуру риторики на более глубоком уровне, чем простое чтение теста. Тексты стратегий являются не самым большим массивом данных, исследователю было бы по силам их все прочитать, однако, по мнению авторов, применение автоматических методов дает лучшие результаты для выявления риторики документа и его прагматического контекста.

В исследовании использовались три разных способа автоматического анализа текста — для идентификации агентов и акторов применялся метод естественной обработки языка; семантические техники — для поиска «актов» через поиск сказуемых, связанных с акторами; машинное обучение позволило проанализировать «сцены» в терминах Бёрка, в рамках которых располагались акторы и их действия. Всего было обнаружено десять тематических групп, которые концептуализировались как «сцены» Бёрка (терроризм, угрозы, права человека, экономическое развитие, энергия и другие). Заземление списка найденных акторов и их действий позволило сфокусированно работать с текстом, причем не просто показать, как темы меняются со временем, но как одни и те же акторы присутствуют в разных тематических группах, или действия, которые первоначально возникли в одной сцене, переносятся в другие. Так, авторы обнаружили, что после атаки 9/11 «акторы» и «акты», которые относились к сцене «терроризма», стали распространяться на другие «сцены», относящиеся, например, к вопросам энергетических ресурсов (Mohr et al., 2013).

Среди главных ограничений работы с большими данными называют доступ к ним (Golder, Macy, 2014). Есть те, кто производит данные, — самые обычные люди, которые оставляют электронные следы. Есть те, кто имеет возможность агрегировать данные и получить к ним доступ. Но самые влиятельные — это те, кто имеет возможность их анализировать. Обойти ограничения можно, создавая специальную инфраструктуру, что, однако, требует больших финансовых вложений[6]. Впрочем, есть и более простой путь — обратиться к анализу данных, доступ к которым открыт для всех. Мы можем предположить, что особую роль новые данные будут играть в тех областях, где нет серьезных ограничений к их доступу. Среди них социология культуры, значительная часть данных для которой можно

6. К примеру, в 2011 году на базе Школы инженерных и прикладных наук Колумбийского университета появился Институт анализа данных, в создании которого большую роль сыграла городская администрация Нью-Йорка. Она предоставила университету 15 миллионов долларов. Институт расположился на 44 000 квадратных футах нового кампуса, были наняты десятки исследователей. Сейчас в Институте функционируют нескольких центров: Центр науки о данных, Центр кибербезопасности, Центр финансовой и бизнес-аналитики, Центр анализа данных о здоровье, Центр новых медиа, Центр «Умные города». Все, что можно отнести к анализу поведения человека с помощью больших данных, сосредоточено в рамках работы Центра новых медиа.

представить в виде текстов, для анализа которых уже сейчас имеются современные инструменты компьютерной науки.

Беспрецедентные возможности наблюдения за поведением людей в реальном времени привлекают ученых, которые не имеют бэкграунда в социальных науках, но обладают достаточными навыками для анализа таких данных. Они нередко считают, что в социальных науках с приходом больших данных и инструментов компьютерной науки должны случиться радикальные перемены, в первую очередь связанные с отменой социальной теории. Социологи, вероятно опасаясь колонизации со стороны инженерных наук, предлагают обновленный вариант социологии. На заключительных страницах обзора мы рассмотрим разные варианты будущего социологии как академической дисциплины.

Версии дисциплинарного будущего

Юрисдикция социологии и новые претенденты

Позиция крайнего эмпиризма представлена в статье аналитика Криса Андерсона, который раньше возглавлял журнал «Wired». В 2008 году он провозгласил «конец теории» и необходимость отказа от научного метода в его прежнем виде.

> Сейчас существует лучший путь. Петабайты позволяют нам сказать: «Хватит с нас корреляций». Мы можем анализировать данные без гипотез о том, какие связи должны в них присутствовать. Мы можем поместить все эти цифры в самые большие компьютеры, какие только известны миру, и позволить статистическому алгоритму найти паттерн там, где его не видит наука… Корреляция заняла место каузальности, и наука может развиваться даже в отсутствие когерентных моделей, унифицированных теорий или любого существующего механического объяснения. Нет никакой причины цепляться за прошлое. (Anderson, 2008, цит. по: Kitchin, 2014: 4)

С этой точки зрения социальные науки должна сменить новая атеоретическая наука о данных. Социологи и политологи должны уступить свое место аналитикам данных, которые не обременены теоретическим багажом социальных наук. Аналитики зачастую убеждены, что можно обойтись без заранее продуманных теорий, моделей или гипотез — алгоритмы могут заставить «данные говорить сами за себя». Если раньше исследователи нужны были для генерации данных, то сейчас в этом нет необходимости. Дисциплинарная компетенция менее важна по сравнению с техническими навыками. Данные смогут дать ответы на какие-либо вопросы только после определенных компьютерных манипуляций, соответственно, необходимой является ученая степень в области компьютерных наук, а не в социологии.

Об угрозе для юрисдикции социологии писали еще до того, как большие данные взорвали Интернет. В 2007 году вышла статья с названием, которое говорит

само за себя: «Наступающий кризис эмпирической социологии» (Burrows, Savage, 2007). Ее авторы были озабочены тем, что коммерческие компании имеют дело с данными, которые мечтают заполучить многие социологи. Коммерческая социология, по их мнению, существует как ответ на рефлексивный характер современного капитализма, которому нужны знания и информация, чтобы извлекать еще больше прибыли. Уже тогда социологи увидели в этом опасность: «Мы были обеспокоены, так как расценили это как еще один гвоздь в крышку гроба академической социологии и ее притязаний на юрисдикцию знания о социальном» (Burrows, Savage, 2014: 2). Раньше методы вносили свой вклад в уникальный характер дисциплины, сейчас данные могут появиться без расчета выборки, проведения интервью или фокус-группы.

Информация о признанных исследователях в области анализа больших данных дает возможность увидеть, насколько серьезны опасения по поводу колонизации социальных дисциплин инженерными науками. Список был получен на основе программ нескольких ключевых конференций, которые проводились в области вычислительной социальной науки[7]. Этот список не претендует на то, чтобы быть исчерпывающим в данной области, однако его достаточно, чтобы оценить количество исследователей помимо социологов. Здесь нужно обратить внимание на область знания, в которой участниками была получена ученая степень: чуть меньше половины — в области социальных наук, другая половина защитила диссертации в области естественных, инженерных и компьютерных наук. Сейчас они аффилированы не только с самыми разными университетскими структурами, но и с индустрией (Facebook, Microsoft). Собственно университетские департаменты также представлены примерно в равных пропорциях: в области социальных наук их чуть меньше — 17, из них 7 — по социологии. За небольшим исключением, если исследователь имеет ученую степень в области технических или естественных наук, то и работать в дальнейшем он будет также в структурах в рамках этих направлений.

Разделение на разные области знания сохраняется и в публикациях. Конечно, в выборе журнала авторы свободнее, чем в выборе места работы. Авторы из нашей выборки часто публиковали статьи в междисциплинарных журналах. При этом в социологических изданиях в основном появляются исследователи из социальных наук. Список основных журналов: «Social Networks» — 17, «American Journal of Sociology» — 7, «Social Forces» — 4, «American Sociological Review», «Journal of Mathematical Sociology», «Social Science Research» — по 3 статьи. Мы не видим ни одного автора из нашего списка, который получил бы техническое образование, на сегодняшний день работает в профильном департаменте и при этом публикуется в социологических журналах. Все авторы без социологического бэкграунда, изучающие социальное поведение, выбирают для публикаций престижные междисци-

7. Список конференций: International Conference on Computational Social Science (Helsinki, June 8–11 2015), Computational Social Science Summit (Northwestern University, May 15–17 2015); 2nd Annual International Conference on Computational Social Science (Northwestern University, June 23–26 2016); Quantifying Science (Tempe, October 1 2015).

плинарные издания — «Nature», «Science», «Plos One», «PNAS», «Scientific Reports». Таким образом, можно говорить о том, что юрисдикция социологии действительно оспаривается со стороны других областей знания. Однако пока это не затрагивает собственно социологические рабочие позиции и журналы, концентрируясь в специальном пространстве, предназначенном для междисциплинарных исследований. В силу этого большая часть социологов может не замечать процесс колонизации или не придавать ему серьезного значения. Но есть несколько исключений, о которых дальше пойдет речь.

Доказательная социальная наука и ее призыв «идти от данных»

Исследователи полагают, что производство знания в социологии должно измениться в силу того, что другие дисциплины также стали использовать данные о социальных транзакциях. Как пишет Кинг, «сейчас социальные науки претерпевают исторически важные изменения, когда их большая часть движется от производства знания, свойственного гуманитаристике, к естественным наукам в том, что касается исследовательского стиля, инфраструктуры, доступности данных, эмпирических методов, содержательного понимания и возможности для быстрого и заметного роста» (King, 2013: 165). О каких изменениях идет речь? Господствующий ныне научный стиль американской социологии сформировался к концу 1970-х годов, и сейчас он доминирует на страницах ведущих социологических журналов. Главное его отличие — это применение опросных инструментов и статистики для проверки заранее сформулированных гипотез. Считается, что в исследованиях американской социологии теория предшествует этапу сбора данных, который направлен на поиск статистической поддержки заранее сформулированных гипотез (McFarland, Lewis, Goldberg, 2015; Pontille, 2003). Изначальная формулировка гипотез при таком стиле имеет огромное значение, ведь нет возможности собрать какие угодно данные. Поскольку сейчас данные появляются не в результате усилий исследователей, соответственно, можно ожидать изменений в принятом порядке действий социологического исследования.

Все больше можно встретить работ, посвященных противопоставлению теоретико-ориентированной науки (theory-driven) исследованиям, которые занимают эмпирицистскую позицию «идти от данных» (data-driven science). В таких исследованиях гипотезы могут возникнуть из характера доступных данных (Kitchin, 2014: 6). Эти работы призывают перестать делать вид, что гипотезы формулируются до того, как исследование было начато и окончено (Goldberg, 2015; McAbee, Landis, Burke, 2017). Во многих случаях гипотезы появляются по мере проведения исследования, но в итоговом тексте исследователь создает иллюзию, что гипотезы появляются на основе всех прочитанных источников и направляют действия исследователя. А. Голдберг справедливо пишет о том, что едва ли найдется исследование, автор которого признается, что пока он искал ответ на один вопрос, нашел ответ на совсем другой. Существующий формат статьи задает логику линейного

изложения, которой следуют в силу принятых норм. Данные должны получить необходимое теоретическое оформление, доказывающее, что гипотезы управляли ходом исследования.

Это довольно изящно продемонстрировал М. Теплицкий, когда сравнил, как изменялись тексты социологов от варианта развернутого доклада на конференции до статьи в научном журнале (Teplitskiy, 2016). Помимо прочего, у него была возможность увидеть, на что чаще всего направлена критика рецензентов, меняют ли они теорию или их замечания в большей степени относятся к анализу данных. Если бы социологические исследования действительно запускались теоретически обоснованным вопросом, Теплицкий вряд ли бы обнаружил, что после процедуры рецензирования главным образом меняется теоретическое обрамление статьи, тогда как анализ данных остается без заметных изменений. Казалось бы, между теорией, исследовательским вопросом, данными и анализом должна существовать более-менее устойчивая связь, соответственно, в случае смены теории должен поменяться и анализ данных. Однако в большинстве социологических статей теория меняется, а анализ остается прежним, что позволяет говорить скорее о теоретическом обрамлении, чем о полноценной опоре на теорию.

Возможно, роль данных и их анализа и раньше была более самостоятельной в исследовании, чем это фиксировалось на риторическом уровне. Главным фактором, почему исследование состоялось, вполне мог быть доступ к данным, а не зазор в теоретическом знании, который и подсказал идею исследования. Но именно сейчас ученые призывают отказаться от Sharking (Secretly Hypothesizing After Results Are Known) и начать следовать Tharking (Transparently Hypothesizing After Results Are Known), то есть перестать скрывать, как осуществлялось исследование (Hollenbeck, Wright, 2016). Важно, что Tharking не является data-mining, когда без всяких идей изучаешь данные и просто получаешь закономерности. Речь идет о появлении гипотез, когда данные обнаруживают новые паттерны, дают новые идеи для рассуждений.

В том, чтобы эмпирико-ориентированная социальная наука стала более легитимной, может способствовать использование новых данных. В силу того, что данные создаются без исследователя, в них обнаруживается большой потенциал именно для индуктивного способа анализа (McAbee, Landis, Burke, 2017). Авторы пишут, что нет необходимости противопоставлять такие исследования дедуктивному способу, скорее нужно стремиться к их большей легитимности. Для их обозначения используется специальное наименование — *доказательная социальная наука* (forensic social science), в которой должны объединиться дедуктивный и индуктивный подходы. Исследователи не должны заниматься проверкой данных на наличие всех возможных связей, они также не должны фокусироваться полностью на проверке гипотез, так как можно упустить неожиданные эмпирические находки. Для того чтобы доказательная социальная наука стала полноценной наукой, которая создает и развивает теории, «исследователи должны работать с данными,

находить важные паттерны, а затем делать шаг назад к построению осмысленных аналитических конструктов» (McFarland, Lewis, Goldberg, 2015).

Социологов не впечатляют одни лишь паттерны, поэтому они готовы внести коррективы в исследовательский стиль социологии, однако не собираются отказываться от необходимости развивать социальную теорию и предлагать социологические объяснения. Новые данные могут быть полезны для обнаружения закономерностей, но главное в социальных науках — это их объяснение. В таком случае большие данные могут дать те самые эмпирические загадки, которые должны присутствовать в исследованиях: «Методы больших данных не являются конечной целью, они только часть движения к объяснительной теории» (Halavais, 2015: 587).

Исследователь может не знать заранее, какой он обнаружит паттерн на основе больших данных, однако чрезвычайно важно, чтобы его исследовательские амбиции диктовали ему не останавливаться на одном только паттерне. Например ,техники тематического моделирования использовались для реконструкции связей между дисциплинами через анализ импорта и экспорта языка друг друга. Связи строились на основе 1 000 000 диссертаций, написанных с 1980 по 2010 год в 157 американских университетах. Авторы обнаружили, что методологические (статистика, математика), технологические (компьютерные науки) и абстрактные тематические категории работают на экспорт — их достижения используются в ряде других дисциплин, тогда как сами эти области замкнуты и редко заимствуют язык других наук. Было также подсчитано количество слов, относящихся к внутреннему и к внешнему языку. Оказалось, что социология со временем демонстрирует заметное снижение доли внутреннего языка и увеличение внешнего. На основе этого авторы сделали вывод о том, что социология является типом науки, которая всегда остается в стадии открытий. Эта стадия характеризуется более заметной ролью внешнего языка. Другие науки также могут опираться на внешний язык, однако их собственный продолжает активно развиваться (Macfarland et al., 2013). Исследование выполнено на основе больших текстовых данных, которые анализируются продвинутыми инструментами. Это прекрасная возможность получить интересные результаты об изменениях языка дисциплин и их связей друг с другом. Результаты могут стать отправной точкой, поводом задаться вопросами, почему доминирует внешний или внутренний язык или почему их соотношение меняется со временем. Как пишет Китчин: «Одно дело — найти паттерн, другое — его объяснить. Это требует глубокого знания социальной теории и контекста. По существу, паттерн — это не конечная точка, а начальная для дополнительного анализа, который почти наверняка потребует новых данных» (Kitchin, 2014: 8).

Существует и более радикальное предложение — развернуть социологию от казуальных объяснений в сторону описаний. М. Сэвидж и Р. Берроуз не просто призывают инкорпорировать большие данные в свои работы, но предлагают заняться исследованиями, в которых будет больше паттернов, чем объяснений (Savage, 2009; Burrows, Savage, 2007). Социологи должны серьезно задуматься о причинах

угасающего интереса широкой публики к собственным исследованиям. Предложение Сэвиджа и Берроуза заключается в отказе от каузальности, поскольку социологии так и не удалось предложить убедительных объяснений. Лучшее, чем социология может сейчас заняться, это делать хорошие описания, используя новые методы и данные. В своих рассуждениях Сэвидж опирается на идеи Э. Эбботта о дисциплинарном проекте, в рамках которого социология могла бы существовать без того, чтобы ставить каузальность на первое место, как это возможно в других дисциплинах:

> Одна из главных причин, почему публика перестала интересоваться социологией, это наше снисходительное отношение к описанию. Публика жаждет описания, но мы слишком презираем этот жанр. Сосредотачиваясь на одной только каузальности, мы отказываем в публикации статьям с чистым описанием, даже если описание выполнено с использованием количественных методов и имеет важные содержательные выводы. В то же время коммерческие фирмы платят миллионы за такую работу, получается, что наше общество фактически «описывается» самым детальным образом частными маркетинговыми компаниями. Но мы, хотя и любим считать, что ответственны за публичное знание об обществе, презираем и описания и методы, которые обычно используются для количественных исследований. Наши социальные индикаторы представляют собой почти случайный набор переменных, пригодных для каузального анализа. (Abbott, 2001: 121)

Для Эбботта социология никогда не будет восприниматься всерьез как наука о социальной жизни, пока она не возьмется за описания. Социология все еще не сделала полный разворот в эту сторону, однако, возможно, большие данные приблизят его. Первый шаг в эту сторону был сделан, когда ученые начали рассуждать об исследовании, которое идет от данных с нелинейной последовательностью шагов, как о легитимном варианте дисциплинарного будущего социологии.

Заключение

Не сложно найти массу примеров полезности больших данных для внешнего мира[8]. Однако можно ли уже говорить о достигнутых успехах в социологии? Есть мнение, что пока мы чаще имеем дело с рассуждениями о применении больших данных в социальных науках, чем действительно с исследованиями, которые строятся на их анализе (Halavais, 2015). Социологи до сих пор чаще критикуют возможности больших данных, чем их используют. Среди десяти самых употребляемых

8. Найденные предсказания могут позволить добиться улучшения не только в коммерческой сфере. Один из самых цитируемых примеров — использование поисковых запросов для информации о реальном распространении заболеваний. Обычно в Европе и США информация о гриппе собирается на основе визитов к врачу, данные публикуются каждую неделю с запаздыванием в 1–2 недели. Поисковые запросы дают возможность отслеживать заболевание быстрее, причем можно заранее получить информацию, которая будет соответствовать реальному поведению (Ginsberg et al., 2009).

ключевых слов в статьях, посвященных большим данным, половина относится к обсуждению вызовов и возможностей их использования — *challenge, revolution, opportunity, value, application, future*[9]. Возможно, стоит согласиться, что содержательный прорыв и важные научные результаты ожидают нас впереди. С другой стороны, именно сейчас принципиально важно обсудить, что может измениться в социологии при переориентации исследователя со сбора данных на постановку вопросов к уже существующим массивам.

Наибольшее внимание исследователей пока сосредоточено на том, чтобы определить отличия новой социальной науки, исследования в которой далеко не всегда носят линейный характер. Гораздо реже обсуждается вопрос об эпистемологическом статусе нового типа данных. В статьях почти по умолчанию считается, что большие данные — подлинные и объективные. Если они не создавались по запросу исследователя, то якобы обладают большей надежностью. Однако стоит помнить, что в случае больших данных «исследователь не только лишен возможности влиять на инструмент, но и нередко не может наблюдать его в действии» (Волков, Скугаревский, Титаев, 2016: 51). В создании новых данных большую роль играют электронные механизмы, которые, заметим, создаются и обслуживаются людьми. В этом смысле как никогда важно продолжить задаваться вопросом, с какими же данными мы имеем дело и о чем они могут нам рассказать. Таким образом, один из необходимых шагов исследования должен заключаться в критической оценке производства данных, что поможет избежать ситуации выявления и описания ложных зависимостей (Там же: 53).

Литература

Волков В., Скугаревский Д., Титаев К. (2016). Проблемы и перспективы исследований на основе Big Data (на примере социологии права) // Социологические исследования. №. 1. С. 48–57.

Сивков Д. (2017). Большие данные в этнографии: вызовы и возможности // Социология науки и технологий. Т. 8. № 1. С. 56–68.

Abbott A. (2001). Time Matters. Chicago: Chicago University Press.

Austin C., Fred K. (2016). The Application of Big Data in Medicine: Current Implications and Future Directions // Journal of Interventional Cardiac Electrophysiology. Vol. 47. № 1. P. 51–59.

Bearman P. (2015). Big Data and Historical Social Science // Big Data & Society Vol. 2. № 2. P. 1–5.

Bail C. A. (2014). The Cultural Environment: Measuring Culture with Big Data // Theory and Society. Vol. 43. № 3. P. 465–524.

Bond R., Fariss C., Jones J., Kramer A., Marlow C., Settle J., Fowler J. (2012). A 61-Million-Person Experiment in Social Influence and Political Mobilization // Nature. Vol. 489. № 7415. P. 295–298.

9. По данным Web of Science, на основе 989 статей.

Bott E. (1955). Urban Families: Conjugal Roles and Social Networks // Human Relations. Vol. 8. № 4. P. 345–384.

boyd d., Crawford K. (2013). Critical Questions for Big Data // Information, Communication & Society. Vol. 15. № 5. P. 37–41.

Burrows R., Savage M. (2007). The Coming Crisis of Empirical Sociology // Sociology. Vol. 41. № 5. P. 885–899.

Burrows R., Savage M. (2014). After the Crisis? Big Data and the Methodological Challenges of Empirical Sociology // Big Data & Society. Vol. 1. № 6. P. 1–7.

Burt R. (2004). Structural Holes and Good Ideas // American Journal of Sociology. Vol. 110. № 2. P. 349–399.

Connelly R., Playford C. J., Gayle V., Dibben C. (2016). The Role of Administrative Data in the Big Data Revolution in Social Science Research // Social Science Research. Vol. 59. P. 1–12.

Centola D. (2010). The Spread of Behavior in an Online Social Network Experiment // Science. Vol. 329. № 5996. P. 1194–1197.

Centola D. (2011). An Experimental Study of Homophily in the Adoption of Health Behavior // Science. Vol. 334. № 6060. P. 1269–1272.

Christakis N. A., Fowler J. H. (2013). Social Contagion Theory: Examining Dynamic Social Networks and Human Behavior // Statistics in Medicine. Vol. 32. № 4. P. 556–577.

Coviello L., Sohn Y., Kramer A., Marlow C., Franceschetti M., Christakis N., Fowler J. (2014). Detecting Emotional Contagion in Massive Social Networks // Plos One. Vol. 9. № 3. P. 1–6.

DiMaggio P., Nag M., Blei D. (2013). Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding // Poetics. Vol. 41. № 6. P. 570–606.

Einav L., Levin J. D. (2013). The Data Revolution and Economic Analysis // *Lerner J., Stern D.* (eds.). Innovation Policy and the Economy. Chicago: University of Chicago Press. P. 1–24.

Evans J. A., Aceves P. (2016). Machine Translation. Mining Text for Social Theory // Annual Review of Sociology. Vol. 42. P. 21–50.

Frizzo-Barker J., Chow-White P. A., Mozafari M., Ha D. (2016). An Empirical Study of the Rise of Big Data in Business Scholarship // International Journal of Information Management. Vol. 36. № 3. P. 403–413.

Goel S., Hofman J. M., Lahaie S., Pennock D. M., Watts D. J. (2010). Predicting Consumer Behavior with Web Search // Proceedings of the National Academy of Sciences of the United States of America. Vol. 107. № 41. P. 17486–17490.

Golder S. A., Macy M. W. (2014). Digital Footprints: Opportunities and Challenges for Online Social Research // Annual Review of Sociology. Vol. 40. P. 129–152.

Goldberg A. (2015). In Defense of Forensic Social Science // Big Data & Society. Vol. 2. № 2. P. 1–3.

Ginsberg J., Mohebbi M. H., Patel R. S., Brammer L., Smolinski M. S., Brilliant L. (2009). Detecting Influenza Epidemics Using Search Engine Query Data // Nature. Vol. 457. № 7232. P. 1012–1014.

Granovetter M. (1973). The Strength of Weak Ties // American Journal of Sociology. Vol. 78. № 6. P. 1360–1380.

Halavais A. (2015). Bigger Sociological Imaginations: Framing Big Social Data Theory and Methods // Information, Communication & Society. Vol. 4462. P. 1–12.

Hollenbeck J. R., Wright P. M. (2016). Harking, Sharking, and Tharking // Journal of Management. Vol. 43. № 1. P. 5-18.

Iliadis A., Russo F. (2016). Critical Data Studies: An Introduction // Big Data & Society. Vol. 3. № 2. P. 1–7.

King G. (2013). Restructuring the Social Sciences: Reflections from Harvard's Institute for Quantitative Social Science // PS: Political Science & Politics. Vol. 47. 1. P. 165–72.

King G., Roberts M. E. (2013). How Censorship in China Allows Government Criticism but Silences Collective Expression // American Political Science Review. Vol. 107. № 2. P. 326–343.

King G. (2009). The Changing Evidence Base of Social Science Research // *King G., Scholzman K., Nie N.* (eds.). The Future of Political Science: 100 Perspectives. New York: Routledge. P. 91–93.

Kitchin R. (2014). Big Data, New Epistemologies and Paradigm Shifts // Big Data & Society. Vol. 1. № 1. P. 1–12.

Kitchin R., McArdle G. (2016). What Makes Big Data, Big Data? Exploring the Ontological Characteristics of 26 Datasets // Big Data & Society. Vol. 3. № 1. P. 1–10.

Koonin S. E., Holland M. J. (2014). The Value of Big Data for Urban Science // *Lane J., Stodden V., Bender S., Nissenbaum H.* (eds). Privacy, Big Data, and the Public Good. Cambridge: Cambridge University Press. P. 137–153.

Lazer D., Pentland A., Adamic L., Aral S., Barabasi A-L., Brewer D., Christakis N., Contractor N., Fowler J., Gutmann M., Jebara T., King G., Macy M., Roy D., Van Alstyne M. (2009). Computational Social Science // Science. Vol. 323. № 5915. P. 721–723.

McAbee S. T., Landis R. S., Burke M. I. (2017). Inductive Reasoning: The Promise of Big Data // Human Resource Management Review. Vol. 27. № 2. P. 277–290.

Manovich L. (2011). Trending: The Promises and the Challenges of Big Social Data // *Gold M. K.* (ed.). Debates in the Digital Humanities. Minneapolis: University of Minnesota Press. P. 469–475.

McFarland D. A., Lewis K., Goldberg A. (2015). Sociology in the Era of Big Data: The Ascent of Forensic Social Science // American Sociologist. Vol. 47. № 1. P. 12–35.

McFarland D. A., Ramage D., Chuang J., Heer J., Manning Ch. D., Jurafsky D. (2013). Differentiating Language Usage through Topic Models // Poetics. Vol. 41. № 6. P. 607–625.

Mohr J. W., Bogdanov P. (2013). Introduction-Topic Models: What They Are and Why They Matter // Poetics. Vol. 41. № 6. P. 545–569.

Mohr J. W., Wagner-Pacifici R., Breiger R. L., Bogdanov P. (2013). Graphing the Grammar of Motives in National Security Strategies: Cultural Interpretation, Automated Text Analysis and the Drama of Global Politics // Poetics. Vol. 41. № 6. P. 670–700.

Moody J., Light R. (2006). A View from Above: The Evolving Sociological Landscape // American Sociologist. Vol. 37. № 2. P. 67–86.

Mützel S. (2015). Facing Big Data: Making Sociology Relevant // Big Data & Society. Vol. 2. № 2. P. 1–4.

Pontille D. (2003). Authorship Practices and Institutional Contexts in Sociology: Elements for a Comparison of the United States and France // Science Technology Human Values. Vol. 28. № 2. P. 217–243.

Peterson R. A., Anand N. (2004). The Production of Culture Perspective // Annual Review of Sociology. Vol. 30. № 1. P. 311–334.

Savage M. (2009). Contemporary Sociology and the Challenge of Descriptive Assemblage // European Journal of Social Theory. Vol. 12. № 1. P. 155–174.

Smith K., Christakis N. A. (2008). Social Networks and Health // Annual Review of Sociology. Vol. 34. P. 405–429.

Teplitskiy M. (2016). Frame Search and Re-search: How Quantitative Sociological Articles Change During Peer Review // American Sociologist. Vol. 47. № 2. P. 264–288.

Wellman B. (1979). The Community Question: The Intimate Networks of East Yorkers // American Journal of Sociology. Vol. 84. № 5. P. 1201–1231.

Zwitter A. (2014). Big Data Ethics // Big Data & Society. Vol. 1. № 2. P. 1–6.

Big Data in Sociology: New Data, New Sociology?

Katerina Guba
PhD in Sociology, Junior Research Fellow, Institute for the Rule of Law, European University at Saint Petersburg
Address: Shpalernaya str., 1, Saint Petersburg, Russian Federation 191187
E-mail: kguba@eu.spb.ru

Recently, we are witnessing an aspiration in the social sciences to collect and analyze the data about human behavior that is being produced with an unprecedented depth and scale. In this article, we discuss how this new data may impact sociology. Big Data has been defined in various ways in literature. Some of the latest works reveal that the key definitional boundary marker is not the volume of data produced, but the traits of velocity and exhaustivity. The differences of this new type of data are that it is not created for research purposes, that it covers the entire population, and that it is produced in real-time. There are two ways to answer the question of key changes in sociology in the era of Big Data. First, new online-data can greatly improve traditional sociological subfields which were prevented from being developed because of a lack of data. Now, there are new results based on online-data which shed light on the causal effect of social influence. Big data can also enable the development of new lines of research because of rapidly-developing computational techniques. This is especially important for those research areas which deal with large bodies of text, and most importantly, new techniques can greatly improve the sociology of culture where empirical research has been less developed when compared with theoretical ideas. Secondly, new data can have an impact on the disciplinary project of sociology. The article ends with the discussion of how Big Data can be used to support data-driven sociology, which differs from mainstream sociology where hypotheses are offered a priori, data is collected, and analyses are conducted to determine the degree to which the hypotheses are supported.

Keywords: Big Data, computational social science, forensic social science, data-driven sociology, network analysis, topic models

References

Abbott A. (2001) *Time Matters*, Chicago: Chicago University Press.
Austin C., Fred K. (2016) The Application of Big Data in Medicine: Current Implications and Future Directions. *Journal of Interventional Cardiac Electrophysiology*, vol. 47, no 1, pp. 51–59.
Bearman P. (2015) Big Data and Historical Social Science. *Big Data & Society*, vol. 2, no 2, pp. 1–5.
Bail C. A. (2014) The Cultural Environment: Measuring Culture with Big Data. *Theory and Society*, vol. 43, no 3, pp. 465–524.
Bond R. M. et al. (2012) A 61-Million-Person Experiment in Social Influence and Political Mobilization. *Nature*, vol. 489, no 7415, pp. 295–298.
Bott E. (1955) Urban Families: Conjugal Roles and Social Networks. *Human Relations*, vol. 8, no 4, pp. 345–384.
boyd d., Crawford K. (2013) Critical Questions for Big Data. *Information, Communication & Society*, vol. 15, no 5, pp. 37–41.
Burrows R., Savage M. (2007) The Coming Crisis of Empirical Sociology. *Sociology*, vol. 41, no 5, pp. 885–899.
Burrows R., Savage M. (2014) After the Crisis? Big Data and the Methodological Challenges of Empirical Sociology. *Big Data & Society*, vol. 1, no 6, pp. 1–7.
Burt R. (2004) Structural Holes and Good Ideas. *American Journal of Sociology*, vol. 110, no 2, pp. 349–399.
Connelly R., Playford C. J., Gayle V., Dibben C. (2016) The Role of Administrative Data in the Big Data Revolution in Social Science Research. *Social Science Research*, vol. 59, pp. 1–12.
Centola D. (2010) The Spread of Behavior in an Online Social Network Experiment. *Science*, vol. 329, no 5996, pp. 1194–1197.
Centola D. (2011) An Experimental Study of Homophily in the Adoption of Health Behavior. *Science*, vol. 334, no 6060, pp. 1269–1272.
Christakis N. A., Fowler J. H. (2013) Social Contagion Theory: Examining Dynamic Social Networks and Human Behavior. *Statistics in Medicine*, vol. 32, no 4, pp. 556–577.
Coviello L., Sohn Y., Kramer A., Marlow C., Franceschetti M., Christakis N., Fowler J. (2014) Detecting Emotional Contagion in Massive Social Networks. *Plos One*, vol. 9, no 3, pp. 1–6.
DiMaggio P., Nag M., Blei D. (2013) Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding. *Poetics*, vol. 41, no 6, pp. 570–606.
Einav L., Levin J. D. (2013) The Data Revolution and Economic Analysis. *Innovation Policy and the Economy* (eds. J. Lerner, S. Stern), Chicago: University of Chicago Press, pp. 1–24.
Evans J. A., Aceves P. (2016) Machine Translation. Mining Text for Social Theory. *Annual Review of Sociology*, vol. 42, pp. 21–50.
Frizzo-Barker J., Chow-White P. A., Mozafari M., Ha D. (2016) An Empirical Study of the Rise of Big Data in Business Scholarship. *International Journal of Information Management*, vol. 36, no 3, pp. 403–413.
Ginsberg J., Mohebbi M. H., Patel R. S., Brammer L., Smolinski M. S., Brilliant L. (2009) Detecting Influenza Epidemics Using Search Engine Query Data. *Nature*, vol. 457, no 7232, pp. 1012–1014.
Goel S., Hofman J. M., Lahaie S., Pennock D. M., Watts D. J. (2010) Predicting Consumer Behavior with Web Search. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no 41, pp. 17486–17490.
Golder S. A., Macy M. W. (2014) Digital Footprints: Opportunities and Challenges for Online Social Research. *Annual Review of Sociology*, vol. 40, pp. 129–152.
Goldberg A. (2015) In Defense of Forensic Social Science. *Big Data & Society*, vol. 2, no 2, pp. 1–3.
Granovetter M. (1973) The Strength of Weak Ties. *American Journal of Sociology*, vol. 78, no 6, pp. 1360–1380.
Halavais A. (2015) Bigger Sociological Imaginations: Framing Big Social Data Theory and Methods. *Information, Communication & Society*, vol. 4462, pp. 1–12.
Hollenbeck J. R., Wright P. M. (2017) Harking, Sharking, and Tharking. *Journal of Management*, vol. 43, no 1, pp. 5–18.

Iliadis A., Russo F. (2016) Critical Data Studies: An Introduction. *Big Data & Society*, vol. 3, no 2, pp. 1–7.
King G. (2013) Restructuring the Social Sciences: Reflections from Harvard's Institute for Quantitative Social Science. *PS: Political Science & Politics*, vol. 47, no 1, pp. 165–172.
King G., Roberts M. E. (2013) How Censorship in China Allows Government Criticism but Silences Collective Expression. *American Political Science Review*, vol. 107, no 2, pp. 326–343.
King G. (2009) The Changing Evidence Base of Social Science Research. *The Future of Political Science: 100 Perspectives* (eds. G. King, K. Scholzman, N. Nie), London: Routledge, pp. 91–93.
Kitchin R. (2014) Big Data, New Epistemologies and Paradigm Shifts. *Big Data & Society*, vol. 1, no 1, pp. 1–12.
Kitchin R., McArdle G. (2016) What Makes Big Data, Big Data? Exploring the Ontological Characteristics of 26 Datasets. *Big Data & Society*, vol. 3, no 1, pp. 1–10.
Koonin S.E., Holland M. J. (2014) The Value of Big Data for Urban Science. *Privacy, Big Data, and the Public Good* (eds. J. Lane, V. Stodden, S. Bender, H. Nissenbaum), Cambridge: Cambridge University Press, pp. 137–153.
Lazer D., Pentland A., Adamic L., Aral S., Barabasi A-L., Brewer D., Christakis N., Contractor N., Fowler J., Gutmann M., Jebara T., King G., Macy M., Roy D., Van Alstyne M. (2009) Computational Social Science. *Science*, vol. 323, no 5915, pp. 721–723.
McAbee S.T., Landis R. S., Burke M. I. (2017) Inductive Reasoning. The Promise of Big Data. *Human Resource Management Review*, vol. 27, no 2, pp. 277–290.
Manovich L. (2011) Trending: The Promises and the Challenges of Big Social Data. *Debates in the Digital Humanities* (ed. M. K. Gold), Minneapolis: University of Minnesota Press, pp. 469–475.
McFarland D. A., Lewis K., Goldberg A. (2015) Sociology in the Era of Big Data: The Ascent of Forensic Social Science. *American Sociologist*, vol. 47, no 1, pp. 12–35.
McFarland D. A., Ramage D., Chuang J., Heer J., Manning Ch. D., Jurafsky D. (2013) Differentiating Language Usage through Topic Models. *Poetics*, vol. 41, no 6, pp. 607–625.
Mohr J. W., Wagner-Pacifici R., Breiger R. L., Bogdanov P. (2013) Graphing the Grammar of Motives in National Security Strategies: Cultural Interpretation, Automated Text Analysis and the Drama of Global Politics. *Poetics*, vol. 41, no 6, pp. 670–700.
Mohr J. W., Bogdanov P. (2013) Introduction-Topic Models: What They Are and Why They Matter. *Poetics*, vol. 41, no 6, pp. 545–569.
Moody, J., Light R. (2006) A View from Above: The Evolving Sociological Landscape. *American Sociologist*, vol. 37, no 2, pp. 67–86.
Mützel, S. (2015) Facing Big Data: Making Sociology Relevant. *Big Data & Society*, vol. 2, no 2, pp. 1–4.
Pontille D. (2003) Authorship Practices and Institutional Contexts in Sociology: Elements for a Comparison of the United States and France. *Science Technology Human Values*, vol. 28, no 2, pp. 217–243.
Peterson, R. A., Anand N. (2004) The Production of Culture Perspective. *Annual Review of Sociology*, vol. 30, no 1, pp. 311–334.
Savage M. (2009) Contemporary Sociology and the Challenge of Descriptive Assemblage. *European Journal of Social Theory*, vol. 12, no 1, pp. 155–174.
Sivkov D. (2017) Bol'shie dannye v jetnografii: vyzovy i vozmozhnosti [Big Data and Ethnography: Challenges and Opportunities]. *Sociology of Science and Technology*, vol. 8, no 1, pp. 56–68.
Smith K., Christakis N. A. (2008) Social Networks and Health. *Annual Review of Sociology*, vol. 34, pp. 405–429.
Teplitskiy M. (2016) Frame Search and Re-search: How Quantitative Sociological Articles Change During Peer Review. *American Sociologist*, vol. 47, no 2, pp. 264–288.
Volkov V., Skugarevsky D., Titaev K. (2016) Problemy i perspektivy issledovanij na osnove Big Data (na primere sociologii prava) [Problems and Prospects of the Studies Based on Big Data (the Case of the Sociology of Law)]. *Sociological Studies*, vol. 1, pp. 48–57.
Wellman B. (1979) The Community Question: The Intimate Networks of East Yorkers. *American Journal of Sociology*, vol. 84, no 5, pp. 1201–1231.
Zwitter A. (2014) Big Data Ethics. *Big Data & Society*, vol. 1, no 2, pp. 1–6.