Pitsch W., Emrich E. The frequency of doping in elite sport: Results of a replication study // International Review for the Sociology of Sport. 2012. Vol. 47. № 5. P. 559–580.
Roberts D.L., St. John F.A.V. (2012). Estimating the prevalence of researcher misconduct: A study of UK academics within biological sciences // PeerJ 2:e562; DOI 10.7717/peerj.562. URL: https://peerj.com/articles/562.pdf (дата обращения: 20.02.2015).
Streiner D.L., Norman J.R., Cairney J. Health Measurement Scales: A Practical Guide to Their Development and Use. Fifth edition. Oxford: Oxford Univ. Press, 2015.
Tian G.-L., Tang M.-L., Wu Q., Liu Y. Poisson and negative binomial item count techniques for surveys with sensitive question // Statistical Methods in Medical Research, 0962280214563345, first published on December 16, 2014 (preprint). URL: http://-smm.sagepub.com/content/early/2014/12/16/0962280214563345.abstract (дата обращения: 22.02.2015).
Wijk N. van, de Leew E., de Bruijn J. The effectiveness of a mixed-mode survey on domestic violence in Curacao: Response and data quality // Field Methods. 2015. Vol. 27. № 1. P. 82–96.
Wolter F., Laier B. The effectiveness of the item count technique in eliciting valid answers to sensitive questions. An evaluation in the context of self-reported delinquency // Survey Research Methods. 2014. Vol. 8. № 3. P. 153–168.
Wolter F., Preisendörfer P. Asking sensitive questions: An evaluation of the randomized response technique versus questioning using individual validation data // Sociological Methods and Research. 2013. Vol. 42. № 3. P. 321–353.

© 2016 г.

В.В. ВОЛКОВ, Д.А. СКУГАРЕВСКИЙ, К.Д. ТИТАЕВ

ПРОБЛЕМЫ И ПЕРСПЕКТИВЫ ИССЛЕДОВАНИЙ НА ОСНОВЕ Big Data (на примере социологии права)

ВОЛКОВ Вадим Викторович – доктор социологических наук, PhD, профессор Европейского университета (volkov@eu.spb.ru); *СКУГАРЕВСКИЙ Дмитрий Анатольевич – Doctoral Candidate, Женевский институт международных исследований и развития, научный сотрудник Европейского университета* (dskougarevskiy@eu.spb.ru); *ТИТАЕВ Кирилл Дмитриевич – ведущий научный сотрудник Европейского университета* (ktitaev@eu.spb.ru), *все – Санкт-Петербург, Россия.*

Аннотация. *Статья характеризует феномен Big Data и потенциал социологии в работе с новыми источниками и типами данных. Описываются параметры "больших данных", приводятся примеры, характеризующие сложности и особенности работы с ними. Демонстрируется, что некоторые данные, с которыми социологи работают давно (государственная и отраслевая статистика) в современном (дезагрегированном) виде, создают для аналитика те же возможности и проблемы, что и "большие данные". Разбирается пример социологической работы с большими данными на примере исследования статистических карточек на подсудимых по уголовным делам. Описывается исследовательская логика, вызываемая к жизни появлением больших массивов данных. В ней на первое место ставится интерпретация смысла уже имеющихся данных. В конце излагается критика "больших данных" и обсужден вопрос о целесообразности использования этого концепта на примере экономики, которая пережила схожую "революцию данных".*

Ключевые слова: *количественные исследования • методология социологического исследования • Big Data • социология права*

Статья подготовлена при поддержке РНФ, проект № 14-18-02219.

Введение. Технологические новации последней четверти века создали ситуацию сбора и хранения в автоматическом или почти автоматическом режиме огромных массивов данных. Биллинговые архивы сотовых компаний позволяют наблюдать количество и регулярность коммуникации абонентов на протяжении многих лет. Данные активности в социальных сетях и поисковых запросов в сети Интернет, материалы видеонаблюдения в общественных местах, архивы личных записей кадровых отделов крупных компаний – все это постепенно становится доступным для обработки и анализа, радикальным образом меняя ситуацию в исследованиях человеческого поведения. Часто такие массивы информации называются "большими данными" (Big Data). Появление новых источников данных, связанное с возникновением новых технических возможностей для сбора и хранения информации, ставит перед социологией два принципиальных вопроса. 1. Может ли социология внести вклад в извлечение нового знания из этих массивов информации или же социологам следует оставаться в рамках традиционных для их науки методов сбора информации (преимущественно опросы, интервью, анализ агрегированной статистики), а работой с новыми массивами должны заниматься специалисты из новых областей, смежных big data (так называемые data scientists)? Если ответ на первый вопрос положителен, то возникает вопрос 2. Как должны трансформироваться методы и приемы социологического исследования или анализа в связи с тем, что социология сталкивается с принципиально новыми типами данных?

Дав положительный ответ на первый вопрос, мы в этой статье подробно остановимся на втором, последовательно показывая, что такое Big Data и в чем их особенность с точки зрения анализа данных, обсудим вопрос о переносе методологии и ограничения работы с такой информацией в смежных областях, разберем некоторые ограничения работы с такими данными.

Проблема Big Data хорошо показывает, что представители эмпирических естественных дисциплин и дисциплин социгуманитарных оказываются в одной лодке. Чтобы сохраниться как области знания сейчас, они должны (вместе) решать схожие проблемы – учиться заново работать с новым типом данных и отвечать на вопросы своей науки с опорой на качественно новую информацию. Одновременно отстаивая право на свое существование у новых отраслей – абстрактных "наук о данных" (вообще) или зарождающихся на волне энтузиазма областей типа "digital humanities". Этот контекст совместного и общего для практически всех дисциплин мы попробуем сделать одним из сквозных в статье.

Статья опирается на наш опыт работы, связанный с социологическим анализом массивов первичной статистики в сфере судебной и правоохранительной деятельности в России (массив из более пяти миллионов карточек на каждого подсудимого в стране за 2009–2013 г.). Такие данные не охватывают все варианты Big Data, но работа с ними требовала тщательного анализа мировой практики работы в этой сфере. Сегодня такие исследования – один из немногих примеров реальных исследований подобного плана на российском материале (см.: [Скугаревский (ред.), 2013; Volkov, 2014; Четверикова, 2014; Волков, 2014]).

Что такое Big Data и как их появление меняет наш мир? Определяя феномен Big Data, часто цитируемый теоретик этой сферы Роб Китчин [Kitchin, 2013: 262] выделяет ряд их характеристик: огромные по объему, содержащие терабайты или петабайты данных; высокоскоростные, создающиеся в реальном времени или почти в реальном времени; различные в своей природе, будучи как структурированными, так и нет; в идеале – исчерпывающие, стремящиеся охватить целиком некоторую популяцию или систему; тонко структурированные по степени детализации, старающиеся быть настолько детальными, насколько это возможно, и позволяющие точно соотносить информацию с конкретным событием; взаимосвязанные, содержащие общие поля, которые позволяют соотносить между собой разные массивы данных; гибкие, обладающие чертами экстенциональности (просто добавлять новые поля) и масштабируемости (могут легко расширяться).

Базы данных, отвечающие этим требованиям, начали собираться с начала двухтысячных годов во многих отраслях человеческой деятельности и тогда же стали осознаваться как принципиально новое явление (см., например, [Laney, 2001]). Примерами служат записи активности пользователей различных систем, данные медицинских обследований, записи камер видеонаблюдения, истории перемещения автомобилей, оснащенных навигаторами, поисковые запросы в сети Интернет и т.д. Со временем количество таких данных росло. Понятно, в силу технических особенностей не все массивы данных соответствовали и соответствуют указанным выше требованиям. Где-то в силу технических причин выгрузка данных в центральное хранилище происходит не в реальном времени, а, скажем, раз в месяц. В каких-то случаях никаких смежных массивов данных представить себе невозможно и, соответственно, соотнесение с другими данными через общие поля оказывается невозможным. Но в целом, обладающие подобными характеристиками базы данных стали неотъемлемой частью нашей жизни.

Главной их особенностью был и остается факт, что первоначальная цель их сбора никак или почти никак не связана с последующим анализом – это данные, собираемые в базы для последующего извлечения сведений об уникальном событии. Они нашли широчайшее применение в криминалистике. С кем созванивался абонент X, когда совершалось преступление? Кто входил в эту дверь с 16-03 до 16-15? Первоначальный дизайн систем управления такими базами был ориентирован преимущественно на то, чтобы иметь возможность относительно быстро и с небольшими трудозатратами отвечать на подобные вопросы. Вопросы, как можно использовать такие базы для получения систематического научного знания, остаются в стороне.

Эта ситуация заставляет говорить о том, что происходит принципиальный эпистемологический сдвиг [Kitchin, 2014]. Следуя за идеями Дж. Грея, одного из ученых в сфере компьютерных наук, Китчин говорит о том, что прошлая (нынешняя) модель, которая именуется “вычислительной” (computational) наукой, опиралась на комплексное моделирование сложных явлений, однако в основе ее лежали выборочные или агрегированные данные. При этом сама модель – гипотетический набор параметров, о которых нужно собрать информацию, в эту модель включаемую, – всегда шла впереди процесса сбора данных. Новая же парадигма, которая именуется “исследовательской” (exploratory) наукой, опирается в первую очередь на существующие массивы данных и учится задавать вопросы имеющимся данным, а не формировать запрос на сбор данных.

Понятно, в естественных науках эта проблема стоит если не острее, то серьезнее. Когда уже есть детальнейшая информация о погоде в любой точке мира в любой момент времени, климатолог, по большому счету, должен думать не о том, как получить данные, а о том, как их анализировать. Однако и в социальных науках все больше и больше сталкиваются с этой же проблемой. Если мы откажемся от убогого взгляда на социологию, сводящуюся к опросам общественного мнения, то увидим, что аналогичная ситуация возникает во многих социологических отраслях. Зачем спрашивать людей о предпочтениях, когда собирается детальнейшая статистика продаж по каждому наименованию продукции? Как такая ситуация должна поменять работу исследователя потребления? Зачем детально расспрашивать о видах религиозной активности, если в каждом храме есть камера, которая фиксирует поведение прихожанина? Понятно, в этом случае за рамками остаются столь важные в социологии вопросы о том, каков субъективный смысл того или иного действия. Но многие задачи, де факто решаемые социологами, могут опираться на принципиально новые данные.

Понятно, что важнейшим условием развития такой “исследовательской”, ориентированной на данные науки является открытость держателей данных. Последние конференции по этой тематике в России и за рубежом демонстрируют высокую заинтересованность держателей данных (коммерческие компании, государственные структуры, операторы каналов связи и т.п.) в аналитической работе с этими дан-

ными. Да, на данный момент существует серьезная ориентация на формирование внутри организаций собственных аналитических структур, которые монополизируют доступ к данным. Но есть основания полагать, что необходимость внешней валидации исследовательских результатов в обозримом будущем заставит компании чаще предоставлять данные внештатным исследователям или публиковать их в открытом доступе. Примером служит компания Yelp, которая собирает отзывы посетителей об организациях (преимущественно, ресторанах). С 2013 г. она поддерживает конкурс Yelp Dataset Challenge, призванный улучшить аналитику на основе этих данных. В рамках конкурса любой исследователь может запросить данные миллионов оценок для анализа, и если его изыскания приведут к получению нового знания, претендовать на награду от компании.

Набирающее силу движение за открытость данных, собираемых государственными органами [World Bank, 2014], и наличие алгоритмов, позволяющих собирать и агрегировать данные, например, об активности пользователей социальных сетей, делают массивы, собираемые корпорациями, не единственным источником знания.

Рождение Big Data вызывает несколько принципиальных изменений и ставит несколько важных проблем. Изменения, в первую очередь, связаны со сменой соотношения между приемами сбора данных и их анализа и изменением подходов к анализу данных. Во-первых, возникает необходимость в техниках анализа, ориентированных на большие, а не выборочные массивы – требуются отсутствующие сейчас в широкой практике инструменты оценки надежности связей, позволяющие исключить случайные зависимости, неизбежно возникающие при применении традиционных подходов к анализу массива из миллионов записей. Во-вторых, возникает потребность в новых технологиях подготовки данных к анализу – от исследователя требуются навыки прикладного программирования, данные часто оказываются в распоряжении аналитика в формате, требующего перевода в годный для анализа вид, систематической (автоматизированной) работы по поиску и удалению ошибок. В-третьих, возникают проблемы эффективной анонимизации данных, которая позволяла бы однозначно идентифицировать каждое наблюдение, но при этом делала невозможным получение доступа к персональным данным. В-четвертых, хотя эта проблема и кажется надуманной в обстановке, когда 50 лет действует закон Мура о том, что число электронных компонентов на кристалле микропроцессора удваивается каждые два года, возникают вопросы по поводу доступа к вычислительным мощностям, необходимым для работы с большими данными.

Однако ключевой проблемой “больших данных” является их смысл и интерпретация. Как собрана информация? Как проводилась фиксация? Какие ошибки инструмент измерения вносит в данные? В обычном исследовании ученый если не полностью контролирует эту стадию, то хотя бы имеет представления о том, что происходит в момент сбора или измерения данных. В случае с Big Data исследователь не только лишен возможности влиять на инструмент, но и нередко не может наблюдать его в действии. А если такая возможность и есть, где гарантия, что инструмент всегда работает одинаково? Так, если на проходной одного из заводов компании нет возможности обойти “вертушку” и каждое пересечение работником ворот фиксируется, то на соседнем заводе той же компании, данные с которого попадают в ту же базу данных, принято не прикладывать пропуск, выходя на перекур, и имеется техническая возможность обойти эту “вертушку”. В результате в базе мы увидим радикальные отличия по интенсивности движения через проходные двух заводов. До тех пор, пока механизмы, собирающие данные, создают и обслуживают люди, мы будем постоянно задавать вопрос: с какими же данными мы на самом деле имеем дело, что они значат? При этом Big Data создают иллюзию, что это абсолютные, конечные, подлинные и объективные данные, которые не нуждаются в дополнительной интерпретации. И именно социология обладает особым потенциалом для деконструкции этой иллюзии.

Расширение пространства Big Data[1]. После того как на теоретическом уровне были поставлены вопросы к "большим данным", становится очевидным, что существует еще один важный тип данных, которые имеют те же проблемы, хотя они и не являются Big Data в строгом смысле этого слова. Речь идет о дезагрегированной статистике, собираемой государственными органами. С советских времен в России собирается большое количество данных. Каждый бухгалтерский баланс каждой компании оседает в налоговых и статистических органах. На каждого подсудимого заполняют статистические карточки с десятками полей. Для каждого преступления полиция фиксирует десятки характеристик. Специфика профессиональных интересов авторов такова, что большая часть примеров в этой части взята из работы правоохранительных органов, социологии права. Имеющиеся сведения позволяют говорить, что в медицине, образовании, социальном обеспечении дела обстоят схожим образом.

Итак, государственные органы собирают огромное количество данных, локализованных до индивида (пациент больницы)/события (преступление) или, по крайней мере, до уровня субъекта учета (школа, суд и т.д.). Дальнейшая судьба этих данных печальна. Перемещаясь наверх, они агрегируются без возможности последующей дезагрегации. То есть на уровне федерации в целом нельзя увидеть статистику преступности глубже, чем по регионам. Понять, обеспечивается ли высокий (низкий) уровень преступности в конкретном регионе "выбросом" одного или нескольких районов, или это общая ситуация – невозможно без специальных усилий. С уровня региона мы видим только условные районы /города, не отдельные субъекты деятельности (больницы, школы, полицейские участки) и уж точно не отдельные случаи. Агрегация идет и на временном уровне – поскольку в центре внимания правоохранителей динамика к аналогичному периоду прошлого года (АППГ) – данные группируются по годам, в лучшем случае – по месяцам. И такая ситуация существует, когда технические возможности для сбора и анализа дезагрегированной статистики с использованием современных статистических методов есть.

При этом во многих государственных органах наблюдается движение в направлении сбора (по крайней мере) дезагрегированных, первичных данных [см.: Титаев, 2011]. Судебная система собирает таким образом статистические карточки на подсудимых, суды обязаны публиковать полные тексты своих решений, что позволяет в автоматическом режиме собирать их с сайтов судов и создавать обширные массивы документов, доступных для обработки (см.: [Криминальная…, 2015]). Есть основания надеяться, что тренд не изменится, и ему будут следовать все новые органы власти и государственные структуры. Аналогична ситуация и в большей части развитых стран [World Bank, 2014].

Однако когда мы начинаем думать над тем, как можно работать с такими данными, мы сталкиваемся с теми же проблемами, что и при работе с Big Data. Как и кем собираются эти данные? Какие существуют стимулы для искажения/некачественного ввода этих данных? Какие инструменты при этом используются, каким образом? Как технологии хранения и передачи данных влияют на их целостность и единообразие? О чем вообще эти данные говорят? Без подробного ответа на эти вопросы любая аналитика с использованием этих данных бессмысленна. И это вопросы, которые часто забываются на волне общего энтузиазма по поводу объема доступных данных. В новой реальности мы вынуждены тратить огромные усилия (раньше мы их тратили на сбор полевых данных) на экспликацию смысла данных и механизма их возникновения.

Соответственно меняется принципиально и общая логика исследования. Если в классическом варианте социологическое исследование строится в логике: "гипоте-

[1] Отдельные положения раздела обсуждались на V международной социологической Грушинской конференции "Большая социология: расширение пространства данных" и в рамках курса "Количественные методы" в Европейском университете в С.-Петербурге.

за – концепты – операционализация – инструмент сбора данных – данные – проверка гипотез" [см., напр., Батыгин, 2008], исследование, построенное на Big Data или их функциональном аналоге, имеет принципиально иную логику: "данные – описание механизма их сбора – интерпретация переменных – гипотезы или вопросы – проверка". Более того, исследование может быть выстроено в жанре поиска закономерностей без связей с теоретическими моделями вообще (методика Data Mining). Тогда модель работы будет выглядеть следующим образом: "данные – поиск связей или автоматическое построение моделей – реконструкция методики сбора данных (для значимых переменных) – интерпретация переменных – интерпретация связей".

Именно социологический взгляд, особенно критическая версия социологических исследований, уделяющая внимание вопросам, каков социальный и физический контекст производства конкретной информации, позволяет одновременно вовлекать массивы "больших данных" в исследования и предупреждать ситуации выявления и описания ложных зависимостей и связей, наличие которых объясняется механизмами сбора данных, не закономерностями реальности, которая стоит за этими данными.

Практическая работа с "большими данными" о судимости. Рассмотрим пример, связанный с правоохранительной статистикой и показывающий сложности и преимущества такого подхода к big data. Используя базу данных в 5 млн. решений судов Российской Федерации, было проанализировано влияние социального статуса подсудимого на решение оправдать, осудить, приговорить к реальному лишению свободы, а также на тяжесть наказания [Волков, 2014]. Анализ проводился методами множественной логистической и линейной регрессии с включением контрольных переменных, отражающих основные юридически значимые характеристики преступления, процесса и обвиняемого. Данные статистической карточки на подсудимого, заполняемые в суде, позволяют учитывать большое количество характеристик: было ли преступление завершено или речь идет только о покушении на него, рассматривалось ли дело в обычном порядке или в связи с тем, что обвиняемый полностью признал вину, в упрощенном режиме (особом порядке судебного разбирательства), который гарантирует осужденному назначение наказания в пределах двух третей от максимально возможного. Из значимых характеристик дела в статистической карточке не отмечаются только факты явки с повинной и помощи потерпевшему непосредственно после совершения преступления – смягчающие обстоятельства, предусмотренные пунктами "и" и "к" части первой статьи 61 Уголовного кодекса, наличие которых, в соответствии с частью первой статьи 62 того же кодекса, гарантирует подсудимому, при отсутствии отягчающих обстоятельств, что наказание не будет превышать двух третей от максимального. Остальные факторы, которые могут оказать формальное воздействие на наказание, представлены в карточке и включены в регрессионные модели.

Сила регрессионных коэффициентов устанавливается путем расчета предельных эффектов. Регрессионный анализ показывает наличие устойчивых различий, связанных с социальным неравенством. Система уголовных репрессий направлена, прежде всего, против маргинальных и низкостатусных слоев населения (безработных, рабочих, заключенных), которые составляют абсолютное большинство подсудимых (более четырех пятых) и наказываются более жестко, чем представители групп с высоким статусом. Кроме того, анализ говорит о конфликте между государством и предпринимателями: более жесткое наказание предпринимателей, чем государственных служащих, за одни и те же преступления. Исследование выявило "феномен студента" – устойчиво более мягкое отношение судов к студентам. Закономерности, выявленные количественным исследованием, объясняются спецификой социального взаимодействия и принятия решений в контексте устоявшегося правоприменения и профессиональных установок судей. Более жесткое отношение к безработным объясняется, прежде всего, юридическими рационализациями судей, ассоциирующих поведение безработных с более высокой общественной опасностью, а также более

слабыми шансами на снисхождение ввиду более слабой интеграции в обществе. Привилегии, получаемые правоохранителями и государственными служащими, могут быть объяснены профессиональным опытом судей и их моральными установками.

Однако простое регрессионное моделирование не позволило бы сделать подобные выводы; данных как таковых оказывается недостаточно. Возникают две проблемы, решение которых очень важно в этом контексте. Если для криминолога или статистика достаточно было бы простых закономерностей, для социолога их явно недостаточно.

Первая: необходимо ответить на вопрос, можно ли вообще опираться на эти данные? Как переменные, которые являются контрольными и объясняющими, вообще попадают в базу данных, кто и как их туда вносит? Какие искажения возникают? Для этого необходимо проведение качественного исследования, которое было сделано в судах и в органах Судебного департамента при Верховном суде (обслуживающего эту базу данных). По итогам исследования стало понятно, что процессуальные характеристики дела скорее всего вводятся с большой точностью, так как они используются органами Судебного департамента и вышестоящих судов для выявления неадекватных решений, принятых судами. Существует логический контроль, который на основании процессуальных и содержательных характеристик дела показывает случаи, в которых решение принято неправильно (размер наказания выходит за допустимые пределы) или неправильно описано в базе данных. Каждый такой сбой вызывает разбирательство с судьей или судебным клерком в зависимости от того, идет ли речь о неверном решении или о его неверном описании в базе данных.

Однако характеристики социального статуса никак не могут влиять на решение с формальной точки зрения. Мотива фиксировать эти данные точно у сотрудников суда, на первый взгляд, нет. Однако качественное исследование показало: данные о социальном статусе оказываются в материалах уголовного дела, которое поступает в суд, кодировка (способ описания) социального статуса в судебных и следственных органах совпадает практически всегда, за исключением редких и очевидных случаев (например, следственная статистика не разделяет работников аппарата судов и судей, а судебная разделяет). Соответственно, для работника суда не составляет труда просто перенести эти данные в обязательное поле статистической карточки, и никаких искажений, кроме обычных сбоев ввода, возникать не должно. Однако как заполняется это поле на следствии? Потребовалось дополнительное исследование. Выяснилось, что российские судьи, руководствуясь частью 3-й статьи 60 Уголовного кодекса, обязаны учитывать при назначении наказания “личность виновного, <…> а также влияние назначенного наказания на исправление осужденного и на условия жизни его семьи.” В качестве “личности осужденного” и “условий жизни его семьи” судьи в первую очередь рассматривают социальный статус, особо большое внимание уделяя занятости, характеру этой занятости и профессиональной позиции. Соответственно, судьи в интервью неоднократно говорили, что они в обязательном порядке требуют, чтобы в уголовном деле содержалось документальное подтверждение фактов, связанных с социальным статусом подсудимого. Более того, в случаях неформальной занятости судьи также следят, чтобы эти факты были отражены в материалах уголовного дела. Они утверждают также, что всегда уточняют эти факты у самого подсудимого. Таким образом, если социальный статус и искажается, то в сторону завышения, потому что свидетельские показания о факте работы считаются достаточным основанием, чтобы считать человека работающим, а подсудимый, как правило, стремится представить себя в максимально выгодном свете.

Итак, первая проблема связана с тем, чтобы проверить адекватность существующих данных. Вторая – понять, какие социальные механизмы стоят за выявленными статистическими закономерностями. В ходе интервью с судьями было выявлено, что отсутствие работы с их точки зрения практически исключает самостоятельное “ис-

правление" подсудимого и практически гарантирует рецидив (насколько обосновано такое мнение для нас в данном случае не важно). Особенно сильно это мнение в случаях имущественных преступлений.

В том, что касается студентов, которые устойчиво получают более мягкое наказание, судьи однозначно обозначили позицию: если человек способен поддерживать статус студента, даже если речь идет о средне-специальном учебном заведении, существует большая вероятность того, что он не принадлежит к криминальной субкультуре. Соответственно, с точки зрения судьи, вероятность рецидива невелика, а преступление скорее всего было случайным эпизодом в биографии человека, его "исправление" возможно без изоляции от общества или за меньший срок.

С тем, что касается более редких статусных категорий (соотношение чиновников и предпринимателей), объяснение по итогам качественных исследований оказалось не столь однозначным, зато гораздо более контр-интуитивным. С одной стороны, судьям было не очень просто комментировать практику вынесения приговоров в отношении предпринимателей и чиновников. Типовому судье такие подсудимые встречаются 5–10 раз за карьеру – не чаще раза в год. Соответственно, такие случаи рассматриваются гораздо более индивидуально. Однако один очень важный факт был озвучен: для государственного служащего факт осуждения означает прекращение карьеры раз и навсегда. Сохранить чиновничью позицию, будучи осужденным даже по делу, никак не связанному с профессиональной деятельностью (ДТП, драка), очень сложно. Практически невозможно последующее возвращение на государственную службу. Поэтому, говорят судьи и эксперты, в отличие от предпринимателя для чиновника сам факт осуждения является очень тяжелым наказанием. Соответственно, назначаемое наказание может быть мягче, или, если оно связано с реальным лишением свободы, короче.

Как видно из приведенного примера, работа с большими данными, "изолированными" от "внешнего мира" и живой социальной реальности, требует для адекватной интерпретации и понимания двух важных усилий: большой социологической работы, во-первых, по контролю механизмов сбора данных, во-вторых, по выявлению социального контекста, делающего возможным их интерпретацию. Без этой сугубо социологической работы полученные выводы могут быть не только необоснованными, но и ложными.

Вместо заключения: почему современные эмпирические исследования избегают термина Big Data. Отдельные апологеты Big Data радикально говорят о "конце теории" с приходом эры больших данных [Anderson, 2008]. Когда доступны все большие массивы данных, предельные издержки по проверке гипотез сокращаются. Благодаря компьютерной революции снизилось и время на их проверку. Зачем строить подлежащую проверке регрессионную модель детерминант экономического роста страны, если можно оценить на компьютере два миллиона ее вариантов [Sala-i-Martin, 1997]?

Когда теоретики экономического роста получили в начале 1990-х доступ к богатым страновым данным (например, о смертности или уровне образования), они встали перед дилеммой, похожей на нынешнюю дилемму у социологов, получивших доступ к Big Data. Да, они могли видеть в данных эмпирические закономерности и описать детерминанты интересующего явления (экономического роста). В итоге экономические журналы середины 1990-х переполнены оценками так называемых growth regressions – регрессий экономического роста, показывающих связь между, например, уровнем образования в стране и сотней других характеристик и экономическим ростом. Однако со временем в экономическую профессию пришло понимание того, что это тупиковое направление развития науки, в итоге журналы перестали принимать статьи, основной вклад которых заключался в оценке growth regression по-новому.

Чтобы объяснить, почему это произошло, необходимо вспомнить четыре базовые операции в статистике, отраженные на рис.

Как и социологи сегодня, экономисты 1990-х гг. пережили значительное улучшение процедуры сбора данных, подлежащих их исследованиям. Если тогда это касалось

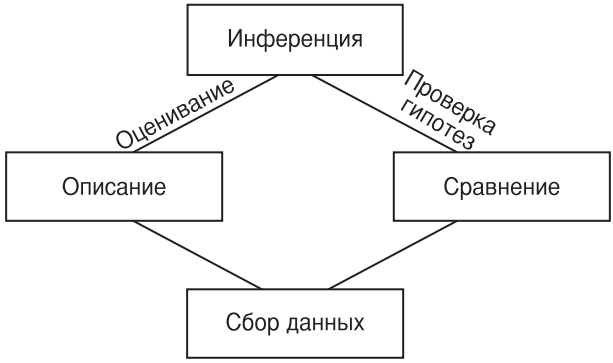

Рис. Базовые операции в статистике
Источник: [Efron, 1982: 342].

страновых данных, то с эрой Big Data стали доступны и дезагрегированные данные. Поскольку в основе любой статистической процедуры лежат данные, улучшение их качества позволило делать более глубокие эмпирические суждения.

Стремительное развитие и доступность статистических пакетов для анализа данных упростили две следующие базовые статистические операции: описание данных (построение описательных статистик) и их сравнение (построение контрастов). В условиях, когда сравнение средней между группами наблюдений, насчитывающих миллионы строк, занимает одну строчку кода, получить базовое понимание структуры данных легко независимо от их размера. Журналы ответили на это повышением требований к плотности информации на квадратный сантиметр страницы: каждый график, каждая таблица должны рассказывать историю, зачастую настолько полную деталей, что описательный подвал к таблице оказывается длиннее самой таблицы.

Несмотря на облегчение трех из четырех операций (сбор данных, описание, сравнение), ни доступность страновых данных для экономики роста, ни Big Data для социологии не повлияли на четвертую операцию – инференцию. Определим инференцию как “получение информации о теоретическом распределении случайных величин (о генеральной совокупности) по выборочным данным” [Цыплаков, 2007: 71]. Упрощая до предела, можно сказать, что речь идет об интерпретации полученных закономерностей – переноса суждения с наличествующих данных на некую реальность. Поскольку любая проверка статистических гипотез строится на подлежащих предположениях, необходимо осознавать сделанные допущения и существующие ограничения.

Первые экономисты, строившие в 1990-х гг. growth regressions на появившихся страновых данных, были воодушевлены богатством подлежащей информации и из-за этого не уделили должного внимания инференции. В итоге оказалось, что множество выявленных эмпирических закономерностей не выдерживают проверки на внешнюю валидность, то есть, не повторяются в других выборках. Поэтому многие результаты, полученные благодаря богатству данных, оказались бесполезны.

Можно возразить, что в случае с Big Data исследователи часто имеют дело с генеральной совокупностью данных, поэтому необходимость применения статистического инструментария, рассчитанного на выборочные данные, вызывает сомнения. Но в реальности даже генеральная совокупность является лишь наблюдаемой выборкой из всех возможных реализаций процесса, генерирующего данные. Например, даже если в современной России собираются данные о всех подсудимых, мы не можем увидеть будущих подсудимых, поскольку они еще не совершили преступлений. Из этого очевидно, что выборочные методы статистического анализа следует применять и при исследованиях генеральных совокупностей.

Эмпирики-экономисты, столкнувшись с проблемой невозможности сделать выводы даже при наличии богатых данных, смогли её решить. В 1990–2000-х гг. экономика стала первой общественной наукой, пережившей "революцию достоверности" [Angrist, Pischke, 2010]. Она заключалась в том, что из четырех базовых операций статистики, применяемых в любом исследовании, главной стала инференция. Львиная доля текстов статей стала посвящаться доказательствам того, что сделанные при инференции предположения разумны и позволяют обнаружить причинно-следственную связь, статистически значимую в генеральной совокупности. Для современного эмпирического исследователя обладать детальным массивом больших данных бесполезно без убедительной инференции в рамках проводимого исследования. На смену growth regressions пришли рандомизированные контролируемые испытания, проводимые в развивающихся / развитых странах [Banerjee et al., 2011] с конкретными испытуемыми. Если раньше для понимания, как, например, микрофинансирование фермеров в развивающихся странах улучшает их благосостояние, экономисты делали оценки на основе рядов данных, предоставленных местным минсельхозом, сейчас ученые самостоятельно формируют эксперименты, где одной случайной группе фермеров дают доступ к микрофинансированию, другой – нет, а потом сравнивается благосостояние контрольной группы и испытуемых [Banerjee et al., 2013].

Однако Big Data все-таки может помочь в инференции эмпирическому исследователю. Богатство больших данных позволяет включать в оцениваемую модель всевозможную информацию. Chan and Tobias [forthcoming] сравнивают такой подход к инференции с киданием в кухонную раковину разных продуктов ("kitchen sink approach") в надежде, что в результате получится что-то дельное. Этот подход ясно обозначает горизонты развития Big Data в общественных науках: исследователи, заинтересованные в выявлении причинно-следственных связей, будут дополнять массивы больших данных в целях их последующей стыковки между собой для получения максимально богатой фоновой информации.

СПИСОК ЛИТЕРАТУРЫ

Батыгин Г.С. Лекции по методологии социологических исследований. М.: РУДН, 2008.

Волков В.В. Влияние социального статуса подсудимого на решение суда // Журнал социологии и социальной антропологии. 2014. № 4. С. 62–85.

Криминальная статистика: механизмы формирования, причины искажения, пути реформирования. Исследовательский отчет / *М. Шклярук, Д. Скугаревский, А. Дмитриева, И. Скифский, И. Бегтин*. СПб.: Институт проблем правоприменения при Европейском университете в Санкт-Петербурге; М.: Норма, 2015.

Скугаревский Д.А. (ред.) Уголовная юстиция в России в 2009 г.: комплексный анализ судебной статистики. СПб.: Институт проблем правоприменения при Европейском университете в Санкт-Петербурге; М.: Статут, 2014.

Титаев К.Д. Как суды принимают решения: исследование влияния внеправовых факторов на российские суды // Экономическая социология. 2011. № 4. С. 122–125.

Цыплаков А. Мини-словарь англоязычных эконометрических терминов. Ч. 1 // Квантиль. 2007. № 3. С. 67–72.

Четверикова И.В. Роль семьи, профессиональной карьеры и пола подсудимых при вынесении приговоров российскими судьями // Журнал социологии и социальной антропологии. 2014. № 4. С. 101–123.

Angrist J., Pischke J.-S. The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics // Journal of Economic Perspectives. 2010. № 24(2). P. 3–30.

Anderson C. The end of theory // Wired magazine. 2008. № 16(7). URL: http: / /archive.wired.com / science /discoveries /magazine /16–07 /pb_theory (дата обращения: 13.05.2015).

Banerjee A.V., Duflo E. Poor economics: A radical rethinking of the way to fight global poverty. Chicago: PublicAffairs, 2011.

Banerjee A., Duflo E., Glennerster R., Kinnan C. The miracle of microfinance? Evidence from a randomized evaluation // CEPR Discussion Papers. 2013. №. 9437. P. 22–53.

Chan J., Tobias J. Priors and posterior computation in linear endogenous variable models with imperfect instruments // Journal of Applied Econometrics, forthcoming. June /July 2015. Vol. 30, Issue 4. P. 650–674.
Efron B. Maximum likelihood and decision theory // Annals of Statistics. 1982. № 10(2). P. 340–356.
Kitchin R. Big Data, new epistemologies and paradigm shifts // Big Data & Society. 2014. April – June. 1–12. URL: http: / /bds.sagepub.com /content /spbds /1 /1 /2053951714528481.full.pdf (дата обращения: 13.05.2015).
Kitchin R. Big data and human geography: Opportunities, challenges and risks // Dialogues in Human Geography. 2013. № 3(3). P. 262–267.
Laney D. 3D Data Management: Controlling Data Volume, Velocity and Variety // Application Delivery Strategies. META Group Report, 06.02.2001. URL: http: / /blogs.gartner.com /doug-laney / files /2012 /01 /ad949–3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (дата обращения: 13.05.2015).
Sala-I-Martin X. I Just Ran Two Million Regressions // American Economic Review: Papers and Proceedings. 1997. № 87(2). № 178–183.
Volkov V. Socioeconomic Status and Sentencing Disparities: Evidence from Russia's Criminal Courts – The European University at St.-Petersburg, The Institute for the Rule of Law. Working paper IRL-01 /2014.
World Bank. Open data for economic growth in Russia. // World Bank Transport & ICT Global Practice Note Series. 2014. July.

© 2016 г.

М.В. РУБЦОВА, Е.А. ВАСИЛЬЕВА

"ДОВЕРИЕ": КОНЦЕПТУАЛИЗАЦИЯ И ОПЕРАЦИОНАЛИЗАЦИЯ ПОНЯТИЯ В КОРПУСНОЙ ЛИНГВИСТИКЕ

РУБЦОВА Мария Владимировна – доктор социологических наук, доцент кафедры социального управления и планирования факультета социологии Санкт-Петербургского государственного университета, Санкт-Петербург, Россия (mariia.rubtcova@gmail.com); *ВАСИЛЬЕВА Елена Александровна – кандидат социологических наук, ведущий научный сотрудник, Академия наук Республики Саха (Якутия), Якутск, Россия* (vasilieva_ea@bk.ru).

Аннотация. *Предложена методика исследования институционально-лексического контекста на основе корпусной лингвистики. Это позволяет концептуализировать и операционализировать понятие "доверие". Методика объединяет количественный и качественный анализ кодированного массива СМИ, с последующей статистической обработкой. В результате исследования выявлено: категория "доверие" используется в контексте политических институтов. При этом личностные качества политиков менее значимы, чем формальный статус и готовность к переговорам.*

Ключевые слова: *доверие • операционализация • концептуализация • институционально-лексический контекст • корпусная лингвистика*

Концептуализация и операционализация понятий – необходимый и сложный элемент программы и методики социологического исследования. Мы предлагаем дополнить имеющиеся подходы достижениями современной лингвистической науки, прежде всего, корпусной лингвистики.