МАТЕРИАЛЫ V СОЦИОЛОГИЧЕСКОЙ ГРУШИНСКОЙ КОНФЕРЕНЦИИ

Американская ассоциация исследователей общественного мнения
(AAPOR)

ОТЧЁТ AAPOR О БОЛЬШИХ ДАННЫХ

12 февраля 2015

Перевод с английского

Москва
2015

УДК 303.432
ББК 60.504
0-88

Джапек, Л., Крёйтер, Ф., Берг, Мю. и др. **Отчёт AAPOR о больших данных: 12 февраля 2015** / Американская ассоциация исследователей общественного мнения; Пер. с англ. Д. Рогозина, А. Ипатовой, Е. Вьюговской; Предисловие Д. Рогозина. М., 2015.

Проектная группа:

Лилли Джапек, зам. директора Бюро статистики Швеции
Фрауке Крёйтер, соруководитель совместной программы по методологии массовых опросов Университета штата Мэриленд и Университета Мангейма
Маркус Берг, Стокгольмский университет
Пол Бимер, Исследовательский международный институт (RTI International)
Пол Декер, Mathematica Policy Research
Клифф Ламп, Школа информации Мичиганского университета
Джулия Лэйн, Американский институт исследований (American Institute for Research)
Кэти О'Нэйл, Исследовательская лаборатория Джонсон (Johnson Research Lab)
Эйб Ашер, HumanGeo Group

Перевод с английского:

Дмитрий Рогозин, заведующий лабораторией ИнСАП РАНХиГС
Анна Ипатова, старший научный сотрудник ИнСАП РАНХиГС
Елена Вьюговская, научный сотрудник ИнСАП РАНХиГС

Перевод выполнен при финансовой и организационной поддержке:
Института социального анализа и прогнозирования РАНХиГС при Президенте РФ
Всероссийского центра изучения общественного мнения (ВЦИОМ)
Фонда содействия изучению общественного мнения (ВЦИОМ)
Автономной некоммерческой организации «Социальная валидация»

Содержание

ПРЕДИСЛОВИЕ К ИЗДАНИЮ НА РУССКОМ ЯЗЫКЕ: ОТ ДОГАДОК К БОЛЬШИМ ДАННЫМ

На зимней социологической школе ВЦИОМ и РАНХиГС 14 февраля 2015 г. рассказывал ребятам, собравшимся в пансионате «Солнечный», о фальсификациях и возможностях перехода от бумаги к планшетам в массовых опросах, важности рассмотрения больших данных для контроля качества и корректировки содержательных выводов. Когда готовился к школе, из Твиттера узнал две потрясающие новости. Буквально накануне, 13 февраля в Гарварде проходил семинар о проблемах фабрикации данных массовых опросов [New frontiers, 2015], а 12 февраля на вебсайте Американской ассоциации исследователей общественного мнения был опубликован отчёт рабочей группы о больших данных [AAPOR report, 2015]. Мы давно свыклись с мыслью о нашем тотальном отставании в методической культуре от англо-саксонского мира. В лучшем случае повторения и пересказы куда более продвинутых зарубежных коллег, в худшем — местечковые попытки сделать что-то новое и оригинальное, зачастую заканчивающееся очередной методической профанацией. Сегодня, в череде значимых методических событий, можно утверждать о своевременности и актуальности поднимаемых нами вопросов.

За 2014 год в России были реализованы две чрезвычайно важные методические инициативы. Во-первых, проведен методический анализ валидности и надёжности масштабного опроса общественного мнения о присоединении Крыма, проведенного двумя полстерскими компаниями — ФОМ и ВЦИОМ. За четыре дня ими был подготовлен инструментарий, проведен общероссийский опрос, охватывающий более 50 тысяч респондентов, и представлен отчёт о результатах. Полученные данные противоречили представлениям протестной группы общества, поэтому в прессе и социальных сетях была развёрнута широкая дискуссия, основной мотив которой состоял в обвинении полстеров в «продажности», «подтасовке данных», «агитационном эффекте опроса». Российские интеллектуалы не могли и не хотели верить в столь фантастические возможности опросной технологии, тем более, когда они привели к неудобному результату — беспрецедентной поддержке россиянами крымской кампании. Нам, группе исследователей, объединенных в неформальное содружество «Методический цех», потребовалось куда больше времени, чтобы провести методическую экспертизу проведенного опроса. За месяц были проанализированы сопутствующие данные массового опроса (параданные), позволяющие выносить суждение об ошибках репрезентации, и проведено когнитивное кодирование случайно отобранных интервью, для оценки ошибок измерения [Вьюговская, 2014; Докторов, 2014; Османов, 2014; Рогозин, 2014; Юдин, 2014]. Это было первое масштабное методическое решение с привлечением пусть не больших данных, но намного превышающих привычные матрицы анкетных ответов. Первый шаг в создании публичного аргумента, основанного на методических экспериментальных планах.

Во-вторых, в течение 2014 года мы вовлеклись в проведение

этнографических наблюдений за работой интервьюеров. Начиная от контроля за деятельностью одной из компаний и заканчивая самостоятельным устройством на работу в другую. Десятки встреч, сотни страниц дневниковых записей и аналитических заметок, обсуждения, споры, наблюдения... За год сформировался не только стройный экспериментальный план, но и были получены первые, отнюдь не вульгарные методические результаты [Вьюговская, Галиева, Рогозин, 2014; Ипатова, 2014; Рогозин, Галиева, 2014]. Опросная технология, основанная на бумажных носителях, под грузом внешних обстоятельств, дала такую трещину, залатать которую уже невозможно ни усиленным контролем, ни призывами к соблюдению этических норм, ни ужесточением внутрикорпоративной этики. Фабрикации и фальсификации сбора эмпирических данных настолько укоренены в индустрии массовых опросов, что воспринимаются их носителями как естественное и непротиворечивое состояние, технологическое решение в ответ на накопившиеся с годами технологические (плохой, перегруженный инструментарий), экономические (низкая оплата труда) и социальные (отсутствие профессиональной подготовки и мотивации к труду) факторы.

Два события международного уровня — семинар в Гарварде и публикация отчёта о больших данных в AAPOR — не просто органически консистентны нашим усилиям, но по целому ряду признаков (тематической наполненности, ценностным заключениям, методическим обоснованием и т.д.) составляют с ними одно целое. Впервые за последние десятилетия можно говорить о совместном, методологически согласованном движении к валидным и надёжным социальным исследованиям, которое мы наблюдаем там, в Соединенных Штатах, и здесь, в России. Поэтому не было никаких сомнений в том, что нужно срочно (пусть и с неминуемыми ошибками и неточностями) осуществить перевод только что вышедшего в свет отчета. Дополнительные усилия по представлению российскому методическому сообществу рамки для дальнейших экспериментальных планов с лихвой окупятся в самом ближайшем будущем. Поскольку нет лучшего способа для понимания окружающей нас действительности, чем детальное наблюдение за попытками и подходами, применяемыми для этого. Обращение социальных исследователей к большим данным — это еще один шаг в сторону внимательного отношения к происходящему, отказу от надуманных и устаревших догматов прошлых лет, формированию новой культуры анализа социальных данных.

Что нового и важного открывает нам отчёт американских коллег? На какие особенности больших данных они обращают внимание? О чём предупреждают и что считают значимыми вызовами для развития методологии социальных наук? Каким образом мы можем принять участие в общем методологическом движении, определяющем и задающем пути развития современного социального и гуманитарного знания?

Связующей рамкой, общей теоретической концепцией для создания методологического нарратива авторами отчёта выбрана общая теория ошибок (см. раздел 5). У нас нет иного пути, нежели внимательное наблюдение за происходящими смещениями и сбоями, восприятие их не как угроз для развития новых технологий, а как возможностей для осмысления

границ их применения, а значит осмысленного включения в корпус современного знания. Если в теории ошибок традиционных опросов выделяются два основных вида: ошибки измерения, связанные с опросным инструментом (анкетой, интервьюером, контекстом разговора и т.д.), и ошибки репрезентации, определяемые особенностями спроектированной и реализованной выборок, то в теории ошибок больших данных авторы отчета предлагают рассматривать три типа: ошибки строки, столбца и ячейки (см. раздел 5.1). Плоская прямоугольная матрица данных, к которой сводится всё многообразие огромных массивов, позволяет вести речь об унификации возникающих ошибок.

Ошибки строки связаны с пропусками элементов массива данных или с их дублированием. Кроме того, в выборку могут попасть элементы, не относящиеся к анализируемому массиву данных. Алгоритмы включения в выборку данных часто изменяются и не контролируются исследователями. Отсюда опасность неустойчивости и спонтанной для исследователя изменчивости массивов становится наиболее критичной для анализа больших данных. Наиболее частая ошибка столбцов данных — некорректное присвоение меток. Автоматическое кодирование может давать сбои, создавая смещения в отобранных переменных. Зачастую речь не идёт о полностью ложных сущностях, фиксируемых в массивах больших данных. Исследователь получает срезы определенной реальности, однако присваивает им ложные значения, за которыми следует ошибочная интерпретация. Ошибки в ячейках авторы подразделяют на ошибки содержания, спецификации и отсутствия данных. Сингулярные отклонения в ячейках могут оказывать значительное воздействие на массив лишь тогда, когда происходит систематический сбой в алгоритмах генерации данных, или производимые данные не соотносятся с интерпретациями их пользователей. Это и есть ошибки спецификации. Пропуски данных, фактически, тождественны основной проблеме массовых опросов — неответам и отказам от участия в интервью. Их значимость увеличивается лишь в случае систематического выпадения тех или иных наблюдений, когда исследователь даже не подозревает, что целые кластеры наблюдений находятся вне зоны его внимания.

Общая теория ошибок больших данных — это универсальный язык описания, позволяющий связать социальные и технические науки, отказаться от восторга по поводу «забавных» распределений или выпадающих за рамки здравого смысла зависимостей и всерьез подойти к новому парадигмальному сдвигу в социальных науках (см. раздел 3.2). Мы уже не моделируем данные, не строим теоретические описания на основе эмпирических наблюдений, не регистрируем происходящее. Информационные потоки, окружающие исследователя, на первый взгляд, хаотичны и несистемны, не поддаются не только пониманию, но и элементарному учёту. Методология больших данных решает эту проблему, предлагая релевантные инструменты генерации, извлечения и анализа данных (см. раздел 5.2). Однако это не снимает необходимости формирования корректной интерпретации. Современный мир погружён в информационные потоки, и наша задача — приспособиться к этой новой

жизни, кардинально отличающейся от привычных книжных описаний.

Концепция общей ошибки в больших данных открывает методологическую перспективу анализа информационных потоков. Авторы отчёта подчёркивают, что большие объемы не гарантируют безошибочности выводов. Напротив, небольшие изменения в алгоритмах сбора или анализа информации могут приводить к значительным видоизменениям и искажением первоначально спроектированного дизайна. Именно поэтому большие данные и традиционные опросы — не конкурирующие, а комплементарные технологии. Обладая своими группами ошибок, врозь они дают куда менее точные и осмысленные данные, нежели совместно. Задача исследователей состоит в разработке комплексного, оптимизированного дизайна исследования, который учитывает особенности изучаемых социальных реалий.

Литература

1 Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Lane, J., O'Neil, C., Usher, A. AAPOR Report on big data: AAPOR big data task force. 2015. 12 February.

2 New Frontiers in Preventing Detecting, and remediating fabrication in survey research / Event by NEAAPOR, Harvard PSR, American Statistical Association. 2015. 13 February. [Regina Faranda, Rita Thissen, Michael Robbins, Alan Zaslavsky, Fritz Scheuren]

3 Вьюговская, Е.В., Галиева, Н.И., Рогозин, Д.М. Этнография «бумажных» квартирных опросов // Мониторинг общественного мнения: экономические и социальные перемены. 2014. № 5. С. 31-55.

4 Вьюговская, Е.В. Призрачная сжатость формулировок вопросов (на материале телефонного опроса о Крыме) // Мониторинг общественного мнения: экономические и социальные перемены. 2014. № 2. С. 57-62.

5 Докторов, Б.З. Заметки об анализе корректности «Крымского опроса» ВЦИОМ и ФОМ // Мониторинг общественного мнения: экономические и социальные перемены. 2014. № 2. С. 26-39.

6 Ипатова, А.А. Насколько разумна наша вера в результаты опросов, или нарушение исследовательской этики в социологических исследованиях // Мониторинг общественного мнения: экономические и социальные перемены. 2014. № 3. С. 26-39.

7 Османов, Т.Э. Особенности построения выборки в опросе о Крыме // Мониторинг общественного мнения: экономические и социальные перемены. 2014. № 2. С. 40-52.

8 Рогозин, Д.М. Насколько корректен телефонный опрос о Крыме: апостериорный анализ ошибок измерения // Мониторинг общественного мнения: экономические и социальные перемены. 2014. № 2. С. 4-25.

9 Рогозин, Д.М., Галиева, Н.И. Методическая (не)устойчивость массового опроса // Полития: Анализ. Хроника. Прогноз. 2014. № 3. С. 169-180.

10 Юдин, Г.Б. Эксперимент под внешним управлением: риторика и репрезентация крымского мегаопроса // Мониторинг общественного мнения: экономические и социальные перемены. 2014. № 2. С. 53-56.

Дмитрий Рогозин
25 февраля 2015

1. Основные выводы

За последние годы мы наблюдаем увеличение объема статистической информации, описывающей различные социальные феномены, основанные на так называемых «больших данных». Данный термин применяется для различных данных, как определено в настоящем отчете, многие из которых характеризуются не только большим объемом, но и вариативностью, высокой скоростью накопления, естественным способом создания и особыми процессами, необходимыми для анализа и статистического вывода. Изменение в становлении новых типов данных, их доступность, способы сбора и распространения – фундаментальны. Это изменение конституирует парадигмальный сдвиг в массовых опросах (survey researches)[1].

В «больших данных» заложен большой потенциал, но для его полного раскрытия требуется решение некоторых фундаментальных проблем. В этом отчете мы предлагаем примеры различных типов больших данных и описываем их возможности для массовых опросов. Мы также описываем обработку больших данных и обсуждаем связанные с ней проблемы. Основные рекомендации представлены ниже.

1 *Опросы и большие данные представляют собой комплементарные, а не конкурирующие источники данных. Различия в подходах должны рассматриваться как преимущества.*

Любое исследование состоит из вопросов, и один из способов ответа на них – утилизация всей доступной информации, связанной как с новым подходом в использовании старых вопросов, так и с актуализацией новых вопросов, которые вызваны доступностью больших данных для исследования. Однако, решения, основанные на больших данных, неминуемо порождают много затруднений, некоторые из которых лучше решаются традиционными опросными методами.

2 *Американская ассоциация исследователей общественного мнения (AAPOR) должна развивать стандарты применения больших данных в массовых опросах, в которых накапливаются большие объёмы информации.*

Значительная проблема заключается в статистически валидном способе обработки больших данных. Одно из общих заблуждений – это уверенность в том, что объём данных может компенсировать любые недостатки данных.

[1] Здесь и далее «survey research» переводится как «массовый опрос». В литературе встречаются другие варианты, например: «опрос общественного мнения», «социальное исследование», «опросная технология» и т.д. Поскольку перед нами документ, близкий к нормативному типу, мы решили не варьировать с переводом терминологии в зависимости от контекста, но всегда придерживаться выбранного, наиболее близкого варианта. — *Прим. перев.*

AAPOR должна развивать стандарты раскрытия и транспарентности больших данных в массовых опросах. Инициатива AAPOR по транспарентности данных[2] — хорошая модель, которая должна распространяться не только на опросы, но и другие источники данных.

3 AAPOR следует начать работу в частном секторе и других профессиональных объединениях по обучению правилам использования больших данных.

Текущий темп внедрения больших данных сам по себе — проблема. Очень сложно сохранять соответствие между традиционным исследованием и развитием в области больших данных. В свете новых технологий исследование очень быстро устаревает. В настоящее время это основной пробел в деятельности членов Ассоциации. AAPOR необходимо привлекать другие профессиональные сообщества, такие как Американская статистическая ассоциация, Ассоциация специалистов по вычислительной технике, для того, чтобы помочь разобраться в этих вопросах и обеспечить подготовку специалистов, входящих и не входящих в AAPOR.

4 AAPOR надлежит информировать *общественность о рисках и преимуществах использования больших данных.*

Большинство пользователей цифровых услуг не знают о том, что данные об их поведении в сети могут повторно использоваться для других целей, как государственного блага, так и для удовлетворения частных интересов. AAPOR должна принимать активное участие в общественных дискуссиях и проводить обучение журналистов, чтобы совершенствовать публичное представление

[2] Инициатива AAPOR по транспарентности данных (AAPOR's transparency initiative) запущена в ноябре 2009 года как движение по раскрытию информации о методологии исследований. Через год Петер Миллер в своём президентском послании AAPOR назвал инициативу по транспарентности важнейшей задачей ассоциации. После этого десятки крупнейших исследовательских организаций ежегодно вступают в сообщество открытой методологической информации. Транспарентность, или открытость методологической информации опросной компании состоит из пяти шагов: (1) распространение информации о важности методологии для качества опросов; (2) следование кодексу профессиональной этики и практики AAPOR; (3) точное понимание различий между транспарентными и нетранспарентными исследованиями; (4) обеспечение сообщества информацией о важности раскрытия данных; (5) предоставление полной информации о своих исследованиях, достаточной для вынесения оценки об их качестве. Для поддержки инициативы от организации требуется выполнение нескольких простых условий: (1) пройти сертификацию по транспарентности данных; (2) подтвердить, что все сотрудники организации, имеющие отношение к раскрытию данных, обладают необходимыми знаниями о важности открытого представления методологических данных и эффективных процедурах раскрытия; (3) предоставление AAPOR информации о базовых процедурах, необходимых для минимального раскрытия методологической информации; (4) предоставление двух описаний исследовательских проектов, отвечающих условиям раскрытия информации; (5) оплата членского взноса; (6) согласие на регулярный мониторинг выполнения соглашений о раскрытии информации со стороны AAPOR. Подробнее см.: <https://www.aapor.org/AAPORKentico/Transparency-Initiative.aspx> [Дата обращения] 23.02.2015. — *Прим. перев.*

данных. AAPOR следует также обновить положения Кодекса профессиональной этики и практики, имеющие отношение к цифровым данным, полученным в опросах. Ассоциация должна сотрудничать с Наблюдательными советами различных институтов (Institutional Review Boards), чтобы облегчить использование таких данных в исследовательских интересах с точки зрения этики.

5 AAPOR *должна содействовать нивелированию барьеров, связанных с различными случаями использования терминологии.*

Эффективное использование больших данных, как правило, зависит от многопрофильного исследовательского коллектива, включающего, например, специалиста в предметной области, научного сотрудника, программиста и системного администратора. В силу междисциплинарного характера больших данных существует множество понятий и терминов, которые определяются по-разному специалистами различных квалификаций. AAPOR должна помочь устранить этот барьер, информируя своих членов о различных способах использования терминологии. Краткие курсы и семинары являются успешным средством, которое AAPOR может использовать для выполнения этой задачи.

6 *AAPOR* необходимо *взять на себя ведущую роль в сотрудничестве с федеральными ведомствами по разработке необходимой инфраструктуры для использования больших данных в массовых опросах.*

Право собственности на данные четко не определено, равно как и не существует четкой правовой основы для сбора и последующего использования больших данных. Возникает потребность в частно-государственном сотрудничестве для обеспечения доступа к данным и их воспроизводимости. Департамент менеджмента и бюджета Администрации Президента США (OMB)[3] принимает активное участие в федеральных исследованиях и разрабатывает рекомендации по их проведению, таким образом, исследования, финансируемые правительством, должны отвечать данным рекомендациям. Важно, что AAPOR работает вместе с федеральными статистическими учреждениями по вопросам изучения больших данных и наращивания потенциала в этой области. Участие AAPOR также предполагает создание или продвижение общедоступных ресурсов для облачных сервисов.

[3] Департамент менеджмента и бюджета Администрации Президента США (The Office of Management and Budget (OMB). Основная задача департамента — содействие и поддержка Президента в формировании и управлении федеральным бюджетом. Кроме того, департамент организует оценочные исследования по эффективности управления, разрабатывает процедуры и регламенты государственной службы, контролирует адекватность организации обратной связи с целью выявления проблем и противоречий в государственном управлении на этапе их зарождения. Официальный вебсайт: http://www.whitehouse.gov/omb [Дата обращения] 23.02.2015. — *Прим. перев.*

2. Предпосылки и цель отчета

Американская ассоциация исследователей общественного мнения (AAPOR) является профессиональной организацией, деятельность которой связана с изучением «общественного мнения», в широком понимании включающего в себя отношения, нормы, ценности и модели поведения. AAPOR постоянно работает над совершенствованием методов и подходов сбора необходимых данных, обучает своих членов, а также политических деятелей, представителей средств массовой информации и общественности в целом, чтобы помочь им более эффективно использовать данные опросов и делать выводы, а также информирует их о новых разработках в этой области. Именно в этом контексте Совет Ассоциации видел острую необходимость написания отчета, резюмируя в нем результаты обсуждений и проблемы, связанные с большими данными, затрагивающие интересы всего сообщества AAPOR.

В связи с тем, что AAPOR предъявляет самые высокие требования к этике исследования общественного мнения, а также способствует развитию методов в области сбора, анализа и интерпретации данных опросов, настоящий отчет учитывает этические вопросы, затрагивающие использование больших данных, и предлагает возможные варианты методов и стандартов работы с большими данными в отдельности или в сочетании с данными опросов.

Существует множество различных норм, предъявляемых к качеству данных и определяющих критерии измерения качества помимо точности. Например, Министерство статистики Канады разработало собственный кодекс норм. Аналогичный представлен в Евростате (Статистической службе Европейского Союза), требования которого обязательны для всех европейских национальных статистических институтов. Европейский кодекс предусматривает следующие критерии качества: актуальность, точность и надежность, своевременность, последовательность и сопоставимость, а также доступность и ясность. Эти параметры также играют важную роль при оценке качества больших данных. Безусловным сдвигом, о котором также упоминается в нашем отчете, стало доказательство того, что большие данные существенно улучшают своевременность проведения исследований и получения выводов. Воздействие на другие критерии качества будет зависеть от источников больших данных и запросов пользователей.

Выработка статистической информации, отвечающей данным критериям качества, теоретически должна опираться на обоснованную методологию, целесообразные статистические процедуры, допустимую нагрузку на респондентов и оправданные затраты (экономическая эффективность). Методология и процедуры будут изменяться в зависимости от используемых источников данных. Общие принципы были учтены рабочей группой в написании данного отчета.

Цели настоящего отчета:

- объяснить, что такое большие данные (Раздел 3);
- описать возможности больших данных (Раздел 4 и 7);
- очертить проблемы, связанные с использованием больших данных (Раздел 5 и 6);
- обсудить возможные решения и исследовательские интересы (Раздел 8).

Отчет является лишь одним из многих документов, которые будут представлять интерес для аудитории AAPOR. Мы хотим указать читателям на стратегическую инициативу, находящуюся в стадии реализации в Американской статистической ассоциации — определить модели для разработки учебных программ, инициировать профессиональное образование, а также взаимодействие с внешними заинтересованными сторонами[4]. В этом же направлении Независимая группа экспертов, по запросу Генерального секретаря ООН[5], разработала конкретные рекомендации по инициированию преобразований, связанных с обработкой и анализом данных в целях устойчивого развития. Экономическая комиссия статистического управления ООН основала рабочую группу в целях разработки ключевых направлений использования больших данных в официальной статистике, а Европейская статистическая система обозначила применение больших данных в стратегическом плане по развитию и финансированию. Кроме того, хотелось бы отметить важность трех отчетов AAPOR: (1) об использовании социальных сетей, (2) мобильных устройств в массовых опросах и (3) неслучайных выборках. Темы отчетов соприкасаются, но имеют свою специфику, поэтому к прочтению рекомендуются все три, чтобы составить полное представление о вопросе. Данные отчетов также представлены на сайте AAPOR[6].

---

[4] http://magazine.amstat.org/blog/2013/06/01/the-asa-and-big-data

[5] http://www.undatarevolution.org

[6] (1) AAPOR. Social media in public opinion research: Report of the AAPOR task force on emerging technologies in public opinion research / J. Murphy, M.W. Link, J.H. Childs, C.L. Tesfaye, E. Dean, M. Stern, J. Pasek, J. Cohen, M. Callegaro, P. Harwood. 2014. 28 May. [Online] <https://www.aapor.org/AAPORKentico/AAPOR_Main/media/MainSiteFiles/AAPOR_Social_Media_Report_FNL.pdf> [Date of access] 23.02.2015;
(2) AAPOR. Mobile technologies for conducting, augmenting and potentially replacing survey: Report of the AAPOR task force on emerging technologies in public opinion research / M.W. Link, J. Murphy, M.F. Schober, T.D. Buskirk, J.H. Childs, C.L. Tesfaye. 2014. 25 Apricl. [Online] <https://www.aapor.org/AAPORKentico/AAPOR_Main/media/MainSiteFiles/REVISED_Mobile_Technology_Report_Final_revised10June14.pdf> [Date of access] 23.02.2015;
(3) AAPOR. Report of the AAPOR task force on non-probability sampling / R. Baker, J.M. Brick, M. Battaglia, M.P. Couper, J.A. Dever, K.J. Gile, R. Tourangeau. 2013. 17 May; Updated 2013. 22 June. [Online] <https://www.aapor.org/AAPORKentico/AAPOR_Main/media/MainSiteFiles/NPS_TF_Report_Final_7_revised_FNL_6_22_13.pdf> [Date of access] 23.02.2015. — *Прим. перев.*

3. Что такое большие данные?

Термин «большие данные» — это своего рода описание больших по объему и разнообразных по составу характеристик, практик, технических приемов, этических проблем и последствий, которые связаны с данными.

Большие данные пришли из естественных наук – физика и астрономия первыми ввели многие из технических средств, которые теперь называют большими данными. Такое оборудование, как Большой андронный коллайдер и квадратная километровая решетка по сути являются огромными хранилищами экзабайтов информации. Способность собирать такой огромный объем данных неизбежно повлекла за собой увеличение возможностей в их обработке и анализе.

Совсем недавно большие источники данных стали обрабатываться с целью нахождения взаимосвязей в экономических и социальных системах, где ранее при помощи опросов, экспериментов и этнографических наблюдений выводились заключения и строились прогнозы. Ниже приведены недавние примеры. Не все из этих примеров могут прямо соответствовать тому, что люди имеют в виду, когда размышляют о больших данных, тем не менее, все они обладают теми характеристиками больших данных, которые приведены ниже (в разделе 3.1).

Пример 1. Онлайн цены

Проекты Массачусетского технологического института Billion Prices Project и PriceStats[7] — академическая инициатива, предлагающая использовать для проведения экономических исследований цены, которые ежедневно собираются с сайтов интернет-магазинов по всему миру. Одной из статистических разработок является оценка инфляции в США. Изменения инфляционных тенденций быстрее отражаются в PriceStats, чем в Индексе потребительских цен[8] (Consumer Price Index – CPI), который публикуется ежемесячно. На рис. 1 приведены совокупные показатели инфляции по месяцам в США с 2008 по 2014 годы; данные, полученные посредством Индекса PriceStats, обозначены оранжевым цветом, посредством расчетов Индекса потребительских цен – голубым.

Сегодня некоторые национальные институты статистики в Европе используют интернет-роботов, которые собирают данные о ценах розничных продавцов с веб-сайтов или считывающих устройств, и предоставляют эти

[7] http://bpp.mit.edu

[8] Consumer Price Index (CPI) — индекс потребительских цен, измеряющий средние изменения в стоимости корзины товаров и услуг. Является главным индикатором инфляции в стране, высчитывается ежемесячно. Индекс готовит Бюро статистики и занятости (Bureau of Labor Statistics) — *Прим. перев.*

данные для Индекса потребительских цен [Norberg et al. 2011, ten Bosch and Windmeijer 2014].

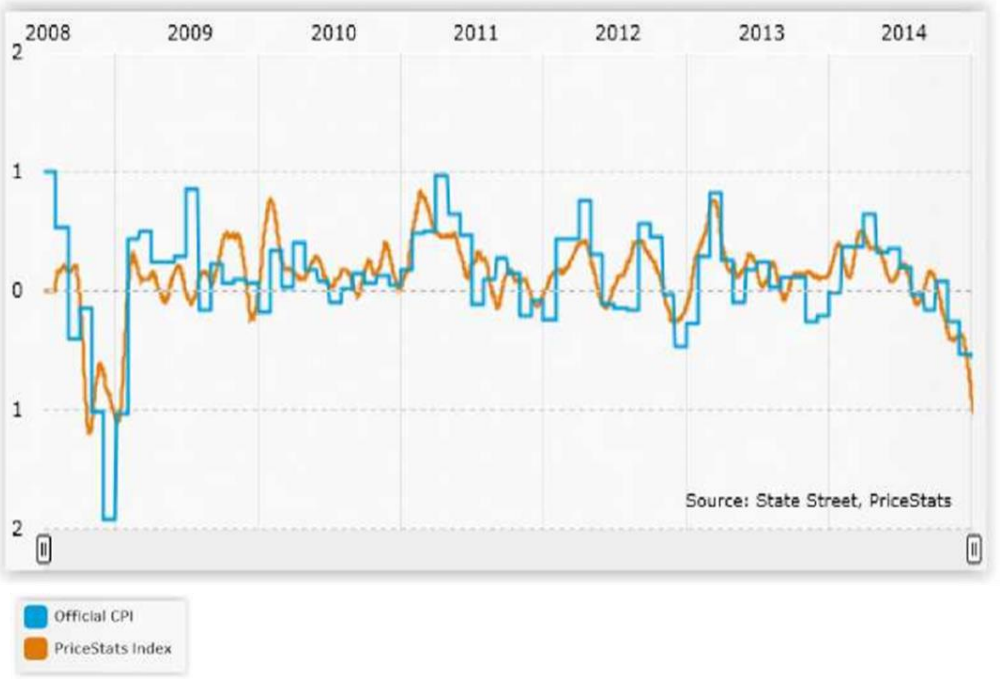

Рисунок 1 - Совокупные показатели инфляции в США, по месяцам, Индекс PriceStats и официальный Индекс потребительских цен (Official CPI).
Данные получены 18 января 2015 года с сайта PriceStats.

Пример 2. Транспорт и инфраструктура

Большие данные могут использоваться для наблюдения за транспортом или нахождения инфраструктурных проблем. Например, на рис. 2 показано количество транспортных средств на улицах в Голландии в течение одного дня. График построен на примерно 100 миллионах записей, полученных посредством дорожных контуров (traffic loops)[9]. При помощи таких данных можно оценить использование инфраструктуры. Подобным образом в

[9] Дорожные контуры (traffic loops) — устанавливаются в местах скопления транспорта (перекрестках, дорожных развязках, пешеходных переходах) и позволяют в режиме реального времени передавать информацию о наличии в точке наблюдения транспортного средства — *Прим. перев.*

Бостоне существует общедоступное приложение для смартфона, разработанное специально для автоматического определения проблем дорожного покрытия[10]. Любой пользователь мобильного приложения может вводить данные о ровности дороги. Согласно веб-сайту, эти данные обеспечивают город информацией в реальном времени, что позволяет устранять проблемы и планировать долгосрочные вложения.

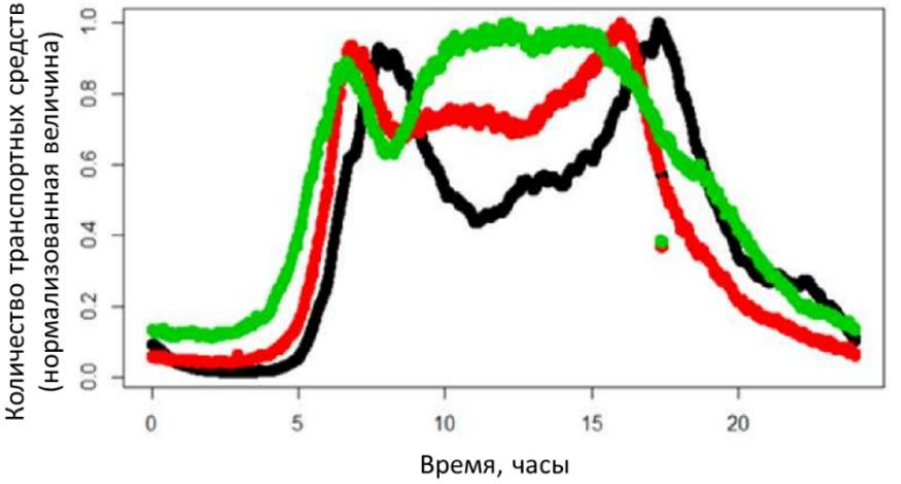

Рисунок 2 - Количество транспортных средств, зафиксированных в Нидерландах 1 декабря 2011 года.
Создано Statistics Netherlands (Daas et al. 2013).
Размер транспортного средства показан разными цветами: черным – малый, красным – средний, зеленым – большой

Пример 3. Сообщения в социальных сетях

Каждый месяц в Статистическом бюро Нидерландов (Statistics Netherlands)[11] рассчитывается Индекс потребительского доверия по данным опросов. Индекс измеряет мнения домохозяйств об их финансовом положении и экономическом климате в целом. Даас и Путс [Daas, Puts, 2014] изучили сообщения в социальных сетях, чтобы понять, могут ли они быть использованы для измерения мнений. Они обнаружили, что корреляция между мнениями в социальной сети (в основном, по данным Facebook) и потребительским доверием очень высока (рис.3).

[10] http://bit.ly/1vrMKHB

[11] Статистическое бюро Нидерландов (Statistics Netherlands) — правительственное учреждение, собирающее официальную статистическую информацию в Нидерландах (Centraal Bureau voor de Statistiek). С 2004 года является полуавтономной неправительственной организацией — *Прим. перев.*

Сообщения в социальных сетях (в данном случае в Твиттере) лежат в основе созданного Мичиганским университетом Индекса потери работы (University of Michigan Social Media Job Loss Index[12]), целью которого является раннее прогнозирование исков о получении страховки по безработице. Прогнозы основаны на факторном анализе сообщений социальных сетей, в которых упоминается потеря работы и связанные с этим последствия [Antenucci et al. 2014].

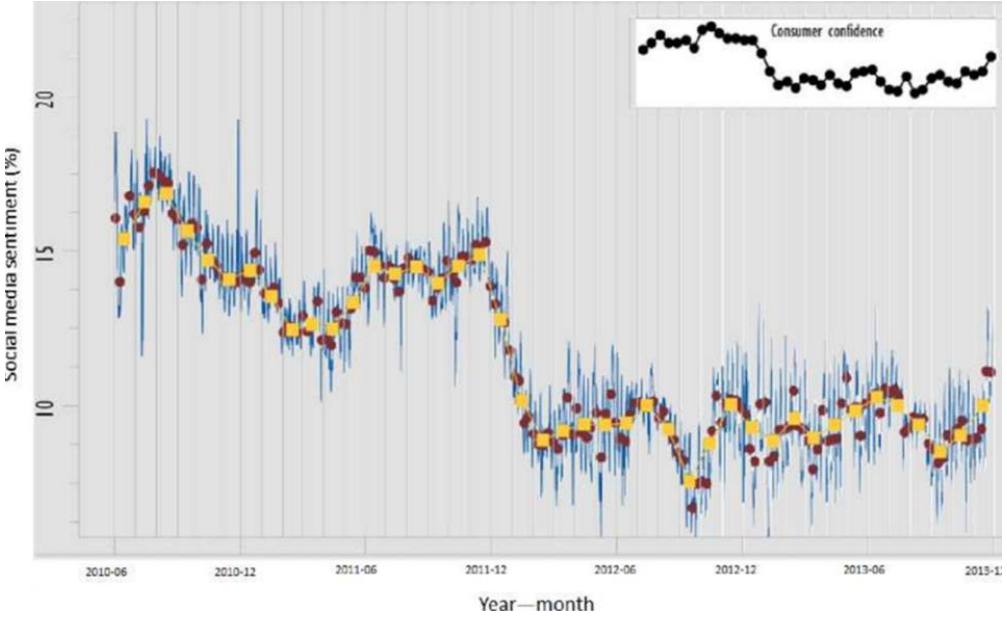

Рисунок 3 - Мнения социальных сетей (по дням, неделям и месяцам) в Нидерландах в июне-ноябре 2013 года. Изменения индекса потребительского доверия за тот же период приведены во вставке (Daas and Puts 2014)

3.1. Характеристики больших данных

Для того, чтобы понять, когда и каким образом большие данные могут выступать в качестве подходящего метода для социологических исследований, необходимо больше узнать о разных характерных особенностях больших данных. Поскольку общепринятого определения

[12] http://bit.ly/1meDoas

больших данных не существует, приведем наиболее распространенное определение из отчета компании Гартнер (Gartner)[13] за 2001 год [Laney 2001, Laney 2012], где описаны некоторые характеристики больших данных:

- **Объем**[14]: Этот пункт относится к огромному количеству данных, доступных для анализа. Большой объем данных связан с постоянно увеличивающимся числом инструментов для сбора данных (например, ресурсы социальных сетей, мобильные приложения, сенсорные устройства), а также с увеличивающейся возможностью хранить и передавать такие данные, связанной с последними усовершенствованиями накопителей информации и компьютерных сетей.
- **Скорость**: Этот пункт относится как к скорости, с которой могут появляться такие случаи сбора данных, так и к давлению, которое возникает ввиду работы с большими потоками данных в реальном времени. В средствах сбора социальной информации новая информация может добавляться в базу данных с разной скоростью: от одного в час до нескольких тысяч эпизодов в секунду.
- **Многообразие** относится к множеству форматов, в которых могут существовать большие данные. Помимо структурированных баз данных, существуют большие потоки неструктурированных документов, изображений, сообщений электронной почты, видео, связей между устройствами и прочей информации, что создает разнородную по своему составу совокупность наблюдений. Одним из следствий такого многообразия данных является то, что их структурирование и сведение требует значительных усилий, и потому эта задача становится основной в работе с большими данными.

Другие исследователи дополнили определение следующими характеристиками: **Изменяемость** (непоследовательность данных во времени), **Достоверность** (возможность безошибочно доверять данным), и **Комплексность** (необходимость стыковать многочисленные источники данных).

---

[13] Гартнер (Gatner) – одна из самых влиятельных компаний в исследовательской и консалтинговой сфере с более чем миллиардным оборотом. В основном специализируется на исследовании рынка информационных технологий. Компания была основана в 1979 году Гидеоном Гарнтером, главный офис компании расположен в Стэмфорде (штат Коннектикут, США). По данным с официального сайта, компания Gatner осуществляет свою деятельность по всему миру, имеет заказчиков в 85 странах. Исследования Gatner регулярно цитируются во многих передовых изданиях. Наиболее известный продукт компании – комплексная методология оценки рынка информационных технологий Magic Quadrant («магический квадрант»). Ежегодно компания организует и участвует более чем в 50 конференциях, тематических форумах, симпозиумах и выставках — *Прим. перев.*

[14] В английском языке для описания основных характеристик больших данных используются три буквы “V” – volume (объем), velocity (скорость), variety (многообразие). — *Прим. перев.*

Мы располагаем разнообразными источниками больших данных, например:

1 данные социальных сетей;
2 персональные данные (например, данные с отслеживающих устройств);
3 сенсорные данные;
4 данные транзакций;
5 административные данные.

Существуют разные мнения на предмет того, должны ли административные данные рассматриваться в качестве больших данных. Административные данные обычно велики по объему, генерируются для разных задач и возникают из административного процесса. Также состав административных данных чаще всего конструируется не исследователями. По этим причинам, а также потому, что существует огромный потенциал в использовании административных данных, мы будем рассматривать их в данном отчете. Необходимо отметить некоторые различия между административными данными и другими видами больших данных. Возможности контроля, которые есть у исследователя, а также потенциальные аналитические мощности варьируются в зависимости от типа источника больших данных. Например, чаще всего у исследователя нет возможности хоть как-то управлять данными, полученными с платформ социальных сетей, потому распутывать текст социальных сетей может быть весьма затруднительно. С другой стороны, для получения административных данных статистическое агентство может заключить партнёрское соглашение с соответствующим государственным органом и влиять на их дизайн. Административные данные более структурированы, определенны и о них больше известно, чем, возможно, обо всех других источниках больших данных.

Большие данные как «полученные» данные

Часто в литературе для специалистов не указывают важный для исследователей общественного мнения факт — большие данные чаще всего являются вторичными данными, первоначально предназначенными для других целей. Это означает, что большие данные как правило изначально не имеют прямого отношения к исследовательским задачам, и лишь повторно используются исследователями для социологических наблюдений. Здесь можно говорить о том, что Шон Тейлор назвал разницей между «полученными» и «созданными» данными [Taylor, 2013]. Он считает, что основное отличие больших данных от других подходов социальных наук заключается в том, что данные не создаются специально посредством вмешательства какого-то исследователя. Когда исследователь общественного

мнения создает инструментарий, появляются уровни контроля и планирования, которых принципиально нет в методиках больших данных. Источники больших данных могут иметь всего несколько переменных, в то время как результаты опросов располагают большим набором переменных, интересующих исследователя. В 2011 году бывший директор Бюро переписи населения США Роберт Гроувз опубликовал в журнале Public Opinion Quarterly и своём директорском блоге статью, где описывает схожие различия между «органическими» и «сконструированными» данными [Groves, 2011a, Groves, 2011b].

В контексте изучения общественного мнения, исследователи могут измерять мнения, получив быстрые ответы на темы, которые могут никогда не возникнуть в источниках больших данных. С другой стороны, «полученные» данные социальных сетей не вызваны реакцией на исследование, то есть возникают естественным образом, так что замеры, не связанные с действиями исследователя, могут более точно репрезентировать настоящее мнение или поведение. «Найденные» данные могут показывать поведение, например, журнал шагов в сетевых шагомерах или ранее упомянутые записи о путешествиях; и эти данные могут оказаться куда более точными, чем те, которые «выспрашиваются» в опросах, когда само озвучивание проблемы может привести к смещениям [Tourangeau et al., 2000].

Обычно под рассмотрение попадают только те данные, которые как-то выделяются, следовательно, само название ***большие*** данные говорит о несконструированной природе данных, которые представляют интерес для исследователей общественного мнения. Например, поскольку данные не были созданы специально для исследования, то во многих случаях отсутствовали и меры по получению информированного согласия, что приводит к этическим вопросам. Кроме того, есть и статистические вопросы относительно репрезентативного характера таких данных. Далее в докладе мы рассмотрим эти серьёзные вопросы более глубоко, и не всегда эта проблема оказывается фатальной для предположения, что большие данные могут быть использованы для проведения социологических исследований (раздел 6).

Данные, которые создаются при работе налоговых систем, социальных программ и других нормативных процедур, также являются «найденными» данными. Они не создаются со специальным расчетом на исследовательский вопрос, а скорее являются побочным эффектом соответствующего административного процесса. Аналогично, параданные (некоторые из их видов) являются побочным результатом процесса сбора социологических данных. Во многих случаях эти административные данные значительны по объёму и так же не структурированы, как и большинство источников больших данных.

3.2. Смена парадигмы

Прежде чем перейти к рассмотрению возможностей и юзабилити больших данных, необходимо разобраться в смене парадигмы, которая происходит в связи с появлением новых источников данных. Эти изменения в парадигме связаны с изменениями многих факторов, которые влияют на измерения человеческого поведения: сущность новых видов данных, их доступность, способы их сбора, распространения и взаимодействия с данными из других источников. Последствия этих изменений для изучения общественного мнения фундаментальны как с точки зрения возможного анализа, так и с точки зрения того, кто может выступать в роли аналитика. Поскольку статистическое сообщество уже миновало период опросов и даже административных данных и дошло до понимания того, как можно извлекать данные из социальных сетей, чтобы замерить общественное мнение, из телефонов — чтобы понять или даже спрогнозировать антиправительственные выступления, из финансовых источников — чтобы изучать экономические колебания, также необходимо отметить, что на сегодняшний день некоторые данные находятся в открытом доступе в готовом для использования виде для каждого, кто хочет сопоставлять замеры и наблюдения или же создавать такую аналитику. Когда данные в интернете легкодоступны, это создает возможности скорее для аналитиков-любителей, чем профессионалов.

Изменение свойств нового типа данных трансформативно. Основные характеристики — скорость, объём, многообразие — и способ, которым эти данные собираются, означают, что новая аналитическая парадигма открыта для статистиков и социологов [Hey et al., 2009]. В классической статистической парадигме исследователи формулировали гипотезу, определяли состав выборки, разрабатывали инструментарий и методику выборочного исследования, а потом анализировали результаты [Groves, 2011a]. Новая парадигма означает, что сегодня возможно собирать данные в цифровом виде, семантически их согласовывать, сводить и коррелировать. Такие сопоставления данных могут вызывать доверие [Halevy et al., 2009, Cukier, Mayer-Schoenberger, 2013] или подозрение (Couper 2013), но в любом случае они позволяют совершать совершенно новые расчёты, большинство из которых были невозможны при использовании только данных опроса. Например, одним из новых видов анализа может быть сбор подробных данных об окружающей индивидов среде посредством сенсорных устройств, Google Earth, видеозаписей, фотографий или финансовых операций. Как вариант, анализ может включать ценную и детальную информацию об уникальных и крайне малочисленных группах людей (посредством данных о микробиомах или протоколов поиска в интернете), или же анализ может основываться на совершенно новых единицах анализа, как, например, компьютерных сетях для людей или предприятий, подключения которых могут фиксироваться новыми типами данных (твитами, разговорами по мобильному телефону, административными записями).

Канеман [Kahneman, 2011] отмечает, что новые способы измерения могут поменять парадигму сами по себе.

Изменение парадигмы также означает изменения в проведении исследований общественного мнения. Смена способов работы с данными, а также необходимых для такой работы навыков частично обуславливается затратами на преобразование данных в информацию, пригодную для использования. По сравнению с опросным миром процесс производства в больших данных сильно отличается. Одним из наиболее очевидных преимуществ больших данных является тот факт, что сбор данных в электронном виде в разы дешевле, чем проведение опросов. Соцопросы по своей сути дороги, поскольку требуют значительных трудовых ресурсов для сбора данных. На этом фоне большие данные, основанные на компьютерном программном обеспечении и электронном сборе данных, могут быть во многом более рентабельны даже при условии предварительной покупки техники и её дальнейшего обслуживания. В то же время, несмотря на относительную дешевизну сбора, большие данные могут вызвать существенные затраты при проверке и обработке (Раздел 5), поскольку для этого требуется перераспределение человеческих ресурсов: вместо разработки дизайна и выборки исследования нужно перейти к структурированию, соединению и контролированию новых видов данных.

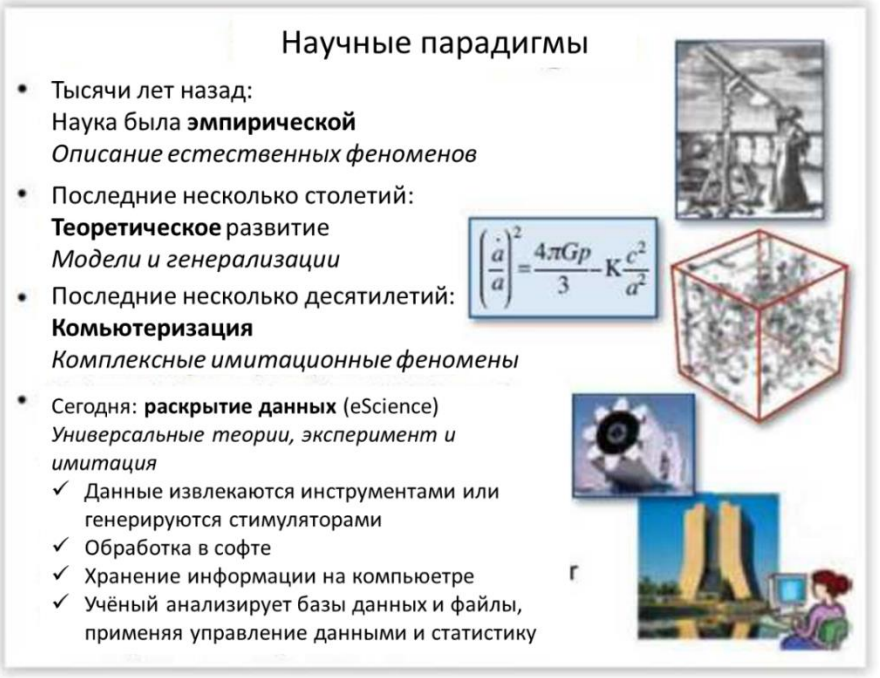

Рисунок 4 - Научные парадигмы [Hey et al., 2009]

Изменения в правах собственности на данные также приводят к изменениям способа распространения данных. Число аналитиков данных — квалифицированных и неквалифицированных — существенно увеличивается. Такой рост может повлечь за собой создание совершенно новых результатов анализа, что показали Слоановский цифровой небесный обзор[15] (Sloan Digital Sky Survey) и проект Polymath[16] [Nielsen, 2012], и что отображено в Четвёртой парадигме Грея[17] (Рис.4) [Hey et al., 2009]. В то же время, это может вызвать снижения качества возможной аналитики, последствий и выводов, сделанных на основе таких данных. Американская ассоциация исследователей общественного мнения (AAPOR) как организация должна будет занять своё место в формировании стандартов использования таких данных в сфере исследований общественного мнения.

Наконец, ажиотаж вокруг смены исследовательской парадигмы должен несколько поутихнуть ввиду осознания того факта, что существующие способы защиты конфиденциальных данных более не жизнеспособны [Karr, Reiter, 2014]. Как только порядок и аналитическая жесткость будут благополучно привнесены в область новых данных, мы должны будем удостовериться, что будущая структура доступа к данным позволит не только заниматься хорошей наукой, но и обеспечит конфиденциальностью тех, кто случайно стал участниками такой науки. Существует большое количество

---

[15] Sloan Digital Sky Survey (**SDSS**) – проект широкомасштабного исследования изображений и спектров звезд и галактик, крупнейшая в мире астрономическая база данных. В рамках проекта были созданы детальные трехмерные карты Вселенной. Официальный сайт: http://www.sdss.org/ [Дата обращения] 23.02.2015. — *Прим. перев.*

[16] Polymath - открытый проект для совместной работы в сети по улучшению оценки в задачах о простых числах. Официальный сайт: http://polymathprojects.org/ [Дата обращения] 23.02.2015. — *Прим. перев.*

[17] Джеймс Николас «Джим» Грей (1944-2007) – известный ученый в области теории вычислительных систем. Его исследовательский интерес был главным образом сосредоточен на базах данных и системах обработки транзакций, где особое внимание уделялось использование компьютера для увеличения продуктивности работы исследователя. Джим Грей работал в компаниях IBM и Microsoft. В 1998 был награжден премией Тьюринга за вклад в развитие баз данных. В последнее время занимался явлением, которому дал название «четвертая парадигма научных исследований», или «наука с использованием большого объема данных». Парадигма Грея встает в один ряд с другими парадигмами науки, по сути, создавая возможность для их интеграции: первой – эмпирической науки, когда описывались явления природы (тысячи лет назад); второй – теоретических умопостроений, для которых использовались модели и обобщения (последние 500 лет), и, наконец, третьей – компьютерного моделирования (последние десятилетия XX века). Четвертая, «вычислительная» парадигма, основанная на использовании большого объема данных, позволяет говорить о создании архива науки, своеобразного расширенного варианта научных публикаций, где наравне с данными, фактами и научными материалами, будут представлены системы коммуникации и вычислительные средства. Эти исследования пока переживают активную фазу роста, приобретая все больше и больше сторонников. Сам Джим Грей, к сожалению, не сможет увидеть архив науки – большой объем постоянно меняющихся и увеличивающихся данных, собранных из разных источников. В январе 2007 он отправился на яхте в однодневное путешествие к Фараллоновым островам и не вернулся. Его тело так и не нашли, и по истечении 5 лет с момента исчезновения Джима Грея признали погибшим в море. С 2007 года компания Microsoft учредила премию имени Джима Грея («Jim Gray eScience Award») для исследователей, которые сделали выдающиеся открытия в области вычислений с большими объемами данных. Личная страница Джима Грея на официальном сайте компании Microsoft: http://research.microsoft.com/en-us/um/people/gray/— *Прим. перев.*

исследований, которые могут быть использованы в качестве отправной точки для такой структуры, но они изолированы друг от друга, находятся в разнородных исследовательских областях, таких, как статистика, информационная безопасность, шифрование. Также есть множество различных практических приложений, в том числе успешная разработка закрытых баз данных с защищенным удаленным доступом. Необходимо соединить знания из этих разных областей и найти способы, при помощи которых новые безграничные наборы данных о человеке могут собираться, объединяться, анализироваться, при этом оставаясь защищенными [Lane et al., 2014]. И здесь Американская ассоциация исследователей общественного мнения (AAPOR) должна будет выступить в том качестве, в котором она выступает сейчас, и включиться в обсуждение.

Выводы

- Термин «большие данные» — это своего рода описание больших по объему и разнообразных по составу характеристик, практик, технических приемов, этических проблем и последствий, которые связаны с данными. Примером источников больших данных могут быть данные о пользовании мобильными телефонами, веб-скрапинг, поисковые запросы, данные сенсорных и считывающих устройств.
- Большие данные иногда называют «найденными» или «органическими» данными, поскольку это своего рода побочный результат процессов, первоочередной задачей которых не является исследование общественного мнения.
- Существуют разные виды больших данных; данные социальных сетей, данные с отслеживающих устройств, данные сенсорных устройств, данные транзакций и административные данные. Возможности контроля, которые есть у исследователя, а также потенциальные аналитические мощности варьируются в зависимости от типа источника больших данных.
- В использовании больших данных есть ряд сложностей и вопросов, например, доступ к данным, этические вопросы, хранение данных и методологические проблемы.
- Большие данные создают новые способы измерения поведения человека. Изменение в сущности новых видов данных, их доступность, способы их сбора, распространения и взаимодействия с данными из других источников фундаментально. Это смена парадигмы для исследований общественного мнения.
- Американская ассоциация исследователей общественного мнения (AAPOR) должна играть фундаментальную роль в определении способов, при помощи которых новые безграничные наборы данных о человеке могут собираться, объединяться, анализироваться, и при этом быть защищены.

4. Почему большие данные важны?

Персональные данные называют «новой нефтью» 21 века [Greenwood et al., 2014], они приносят существенную пользу политике, обществу и исследованиям общественного мнения. Детализированные данные о людях могут быть использованы руководством для снижения уровня преступности, улучшения качества предоставления медицинских услуг, для лучшего управления городами [Keller et al., 2012]. Общество также получает пользу от этих данных. Последние работы показывают, что предприятия, опирающиеся на данные, на 5% производительнее и на 6% доходнее своих конкурентов [Brynjolfsson et al., 2011; McAfee, Brynjolfsson, 2012]. Использование данных большого объема, высокой скорости и уровня разнообразия потенциально может позволить исследователям общественного мнения увеличить масштабы работ по сбору данных, и в то же время уменьшить погрешность, а также денежные и временные затраты [Murphy et al., 2014].

Ценность больших данных для каждой из этих групп (высшее руководство, предприятия и исследователи), баланс выгод и затрат, включая риски использования новых информационных активов, различается, поскольку различается метод расчёта для каждой группы. Польза больших данных для высшего руководства уже давно и хорошо изложена [Lohr, 2012; Koonin, Holland, 2014]. Белый дом отметил, что «технологии больших данных улучшили почти все услуги, которые оказывает государственный сектор» [Executive Office of the President, 2014], но затраты на осуществление этих изменений нетривиальны. Как было отмечено ранее, даже при условии того, что сам сбор данных достаточно недорогой, стоимость контроля, отбора, стандартизации, совмещения и использования данных нового типа может быть существенна (Раздел 5). Зачастую у федеральных, государственных и местных органов нет собственных возможностей для проведения такого анализа [Pardo, 2014], и в результате им приходится открывать доступ к данным для консультантов или научного сообщества, включая разработку протокола и условий доступа. Федеральное правительство, органы штатов и некоторые органы местного самоуправления для решения этих задач вводят должность управляющего директора по данным [Griffin, 2008; Pardo, 2014].

Существуют также и значительные риски, связанные с переходом от традиционных к новым методам сбора данных, одним из которых является нерациональное использование ресурсов. Например, чрезмерное доверие к данным Твиттера при оценке распределения ресурсов для ликвидации последствий урагана может привести к ошибочному перераспределению ресурсов в сторону молодых, умеющих пользоваться Интернетом людей с мобильными телефонами, и оставить пожилых или бедных людей без помощи [Shelton et al., 2014]. Впрочем, все подходы к сбору данных обладают схожими рисками. Слабая методология опросов привела к тому, что журнал Literary Digest неправильно предсказал результаты президентских выборов в 1936 году [Squire, 1988]. Неадекватное понимание вопросов покрытия и

качества опросного инструмента, а также отсутствие контрольных групп тормозили использование административных записей, что особенно касалось записей о преступлениях, на основании которых делались выводы о роли политики смертной казни в сокращении уровня преступности [Donohue, Wolfers, 2006; Levitt, Miles, 2006]. Так же как и в случае с традиционными методами сбора информации, чтобы распознать риски, связанные с новыми подходами, необходимо их документировать (Раздел 5). Еще один риск — отчуждение тех людей, данные о которых собираются. В некоторых сферах не существует чётких правил или рекомендаций о том, как обеспечить конфиденциальность в этом новом мире, где в публичном и частном секторе генерируются доступные для анализа данные [Ohm, 2010; Executive Office of the President, 2014$ Strandburg, 2014]. Аналогично, не существует таких операторов баз данных или депозитариев, которым можно было бы доверить защиту и конфиденциальность [Lane, Stodden, 2013].

Использование больших данных в исследовательских целях также приносит значительную пользу обществу. Товарная продукция может быть эффективно перенаправлена к целевым потребителям, медицинская помощь — лучше спланирована, налогоплательщики могут платить меньше за государственные услуги [Lohr, 2012]. Торговые фирмы также могут получить выгоду от снижения затрат и увеличения производительности [Brynjolfsson et al., 2011; Tambe, Hitt, 2012]. Поставленная задача имеет риски, которые не до конца осознаны и подсчитаны [Barocas, Nissenbaun, 2014]. Улучшенные данные о преступности могут помочь полиции лучше распределять ресурсы, но также могут привести и к усилению национальных конфликтов [Gelman et al., 2007]. Большее количество данных о возможных террористах, как в случае с организаторами взрывов в Бостоне, может способствовать быстрому установлению личности, но также может привести к тому, что невинные жители будут ошибочно идентифицированы как террористы [Tapia et al., 2014].

Интеллектуальный анализ персональных данных может улучшить социальное обеспечение, снизить стоимость поиска решений, уменьшить число случаев экономической несостоятельности; в то же время он может стать источником убытков, экономического неравенства и разбалансирования мощностей между теми, кто обладает данными и теми, чьими данными владеют. Так, компания может снизить свои складские расходы путём интеллектуального анализа данных о поведении многих отдельных потребителей; тем не менее, инфраструктура, необходимая для проведения такого анализа, потребует значительных вложений, и, если проводимый анализ приведёт к раскрытию личной информации клиентов, такие вложения могут иметь негативные последствия. Кроме того, потребитель может получить пользу от предоставления своих данных в огромную базу данных индивидуальных предпочтений (например, поделиться своими музыкальными предпочтениями с онлайн-продавцом и получить в ответ целевые рекомендации о новой музыке, которую можно послушать);

этот же самый потребитель, потеряв контроль над этими данными, может пострадать от кражи личных сведений, ценовой дискриминации и стигматизации, связанной с тем, что стороны, для которых эта информация не предназначалась, могут многое узнать о потребителе [Acquisti 2014, p. 98].

Польза для исследователей *общественного мнения* потенциально огромна. К новому способу сбора данных относятся как к «четвертой парадигме» в науке [Hey et al., 2009], и необходимость соединения социальной и компьютерной наук в виде анализа больших данных признается большинством профессиональных ассоциаций [Schenker et al., 2013]. Одним из очевидных преимуществ является то, что это дает исследователю набор новых аналитических инструментов. В дополнение к тщательному сбору данных, основанному на гипотезах, новые данные обладают, как это отметил Роберт Гроувз [Groves, 2011a], четырьмя особенными качествами, которые необходимо добавить к исследовательскому мировоззрению: (1) стремятся измерить поведение, а не внутренние установки, такие как отношения и убеждения; (2) предлагают записи о явлениях в псевдореальном времени; они очень темпорально раздроблены; (3) основываются на нескольких переменных, у многих есть только какой-то идентификатор и одна переменная (например, текст твита, координата GPS); и (4) редко предлагают четко определённое покрытие большого населения (мы не знаем, кого нет на Facebook, Twitter, поиске Google). Развитие статистических приемов, которые используют богатство данных, но избегают логических выводов, становится критическим [Varian, 2014], так же как и комбинирование источников данных (Раздел 7).

Что интересно, новые данные могут изменить то, как исследователи думают о поведении. Например, они позволяют собирать информацию обо всей окружающей субъекта среде, при этом обеспечивая возможности для понимания эффектов комплексного влияния окружающей среды на человека. Кроме того, некоторые из источников больших данных позволяют исследователям изучать области больших отклонений так, как это нельзя сделать с малыми данными (допуская, что источники данных не страдают от самовыбора). Такие выбросы кривой распределения обычно касаются самых интересных слоев исследуемого населения, которые сложно охватить; учитывая стоимость медицинского обслуживания для небольшого числа заболевших людей [Stanton, 2006], или экономическую активность и трудовые отношения внутри маленьких компаний [Jovanovic, 1982; Evans 1987].

Выводы

- Польза от использования больших данных для улучшения государственных услуг очевидна, тем не менее, затраты и риски на осуществление этих улучшений нетривиальны.
- Большие данные предоставляют совершенно новые способы измерения человеческого поведения в псевдореальном времени. Несмотря на это, измерения поведения часто очень нестабильны.
- Большие данные предоставляют возможность изучать области больших отклонений, ранее недоступные из-за статистических погрешностей.

5. Процесс сбора больших данных и проблема качества

Огромное количество многофакторных и неструктурированных массивов в больших данных привносит одновременно новые возможности и угрозы для анализа. Многие проблемы с большими данными хорошо изучены, что отражено в предыдущих главах (см. Раздел 3, 4). Большие данные зачастую выборочны, неполны и содержат ошибки. Перечислим эти новые ошибки.

Большие данные обычно агрегируются из различных источников, по разным временным основаниям и собираются в некоторые наборы данных. Для этого требуется связывать между собой записи, преобразовывать их, создавать новые переменные, одновременно документируя все преобразования данных. Кроме того, требуется описывать возможные ошибки, возникающий шум и случаи снижения надежности и валидности данных, возникновения смещений. Анализ больших данных с весьма важным условием сохранения статистически валидных выводов весьма проблематично.

Основная проблема слишком большой скорости изменения больших данных — это неконтролируемость инструментов и методов, по которым собираются данные, что ставит под сомнение валидность и надежность, их пригодность для научного анализа. Скорее, как описано в разделе 3, отыскиваемые большие данные часто являются побочными продуктами, косвенными данными, генерируемыми вовсе не для задач анализа. Нельзя игнорировать и риск несоответствия данных, например, автоматизированные системы могут быть специально созданы для генерирования контента. Следовательно, большие данные зачастую имеют весьма слабую связь с качественными, проверенными сведениями. Ответственность аналитиков — чётко осознавать и описывать ограничения данных и предпринимать необходимые меры для ограничения возможных эффектов и ошибок.

Широко известен пример провала Гугловского прогноза заболевания гриппом (Google Flu Trends), связанного с ошибками больших данных. Прогноз основан на анализе поисковых запросов о симптомах гриппа, средств защиты, позволяющей «почти в реальном времени» оценивать вирусную активность в США и еще 24 странах по всему миру. По сравнению с данными Центров контроля и профилактики заболеваний (Centers for Disease Control and Prevention, CDC)[18], прогнозы Гугла о заболеваемости

[18] Центр контроля и профилактики заболеваний Соединенных Штатов (United States Centers for Disease Control and Prevention, US CDC) — крупнейшее федеральное агентство, образованное 1 июля 1946 года в Атланте, штат Джорджиа. Первоначально созданный для контроля распространения малярии на территории США, в настоящий момент агентство вовлечено в десятки федеральных программ по диагностике и предотвращению инфекционных заболеваний, пищевых отравлений, заболеваний, связанных с плохой экологией, безопасностью для здоровья рабочих мест. В центре разработаны собственные опросные

гриппом были удивительно точны с 2009 по 2011 год включительно. Однако в 2012 и 2013 годах Гугловский прогноз гриппа предсказал почти двойной рост обращений к врачам с симптомами гриппа, по отношению к прогнозу CDC [Butler, 2013]. Лазер с коллегами [Lazer et al., 2014] зафиксировали два типа ошибок: в завышенной оценке больших данных и построении алгоритма, оценивающего динамику. Последнее произошло тогда, когда исследователи больших данных понадеялись, что объем данных может компенсировать любые отклонения и дефициты, что позволяет отказать от традиционных научных и аналитических процедур. Как отмечают Лазер с коллегами [Lazer et al., 2014, p. 2], обнаруженная завышенная оценка значимости больших данных указывает на то, что «большое количество данных не должно приводить к игнорированию фундаментальных вопросов измерения и конструированию надежности и валидности...».

Хотя объяснений произошедшего множество, факт остаётся фактом — Гугловский прогноз заболевания гриппом был слишком завышен и дал значимые отклонения с 100 по 108 неделю, начиная с июля 2012 года. Лазер с коллегами [Lazer et al., 2014] также указывают на ошибки в построении динамических рядов, когда генерируемые машиной данные модифицировались таким образом, что построенные ранее точные прогнозы начали давать сбой. Например, когда пользователь составлял запрос по ключевым словам — «лихорадка» и «кашель», Гугл рекомендовал перейти к результатам запросов по симптомам гриппа и лечения, то есть автоматически генерировались данные, впоследствии входившие в расчёт прогноза. Соответственно, в результате изменения алгоритмов расчета, связанные с гриппом запросы генерировались автоматически. В массовых опросах это похоже на смещения, создаваемые интервьюером, когда при наличии кашля респонденту предлагается отметить, что у него, возможно, грипп.

Проблемы динамических алгоритмов не ограничены Гуглом. Такие платформы, как Твиттер и Фейсбук, также часто изменяют и модифицируют конфигурацию пользовательских запросов. Урок, который следует извлечь из истории с гугловским прогнозом заболевания гриппом, состоит в том, что успешный анализ больших данных сегодня может привести к ошибкам завтра. Все эти платформы периодически изменяют алгоритмы расчёта, что приводит к ошибкам в любых долгосрочных исследованиях. Рекомендательные сервисы еще больше усугубляют последствия от подобных инициатив, и беда в том, что оценить их практически невозможно. Кроме того, нам неизвестны другие ошибки, которые могут иметь значимое влияние на гугловские прогнозы. Например, характеристики интернет-аудитории могут вызывать сильные смещения, связанные с генерализацией вывода. Социально-демографический состав пользователей интернета может сильно

системы, позволяющие с высокой точностью прогнозировать распространение заболеваний. Официальный веб-сайт: http://www.cdc.gov — *Прим. перев.*

отличаться от социально-демографических характеристик тех, кто болеет гриппом [Thompson et al. 2006]. Это означает, что люди, имеющие высокий риск заразиться гриппом, и люди, делающие запросы в гугле, не совпадают. Это демонстрирует только одну из проблем репрезентативности, которая часто преследует анализ больших данных. В целом, вопрос заключается в том, что надежность алгоритмов (публично) не проверяется, поскольку чаще всего они являются чьей-то собственностью. Особенность Гугловского прогноза заболевания гриппом в том, что он публично провалился. Из того, что мы видели, несостоятельность большинства моделей проявляется негласно, и этого никто не замечает.

Неточность данных показывает только одну часть проблем для аналитика больших данных. Другие проблемы возникают исключительно как результат огромного размера и многомерности данных. Фань с коллегами [Fan et al., 2014] определяют три типа проблем, а именно: (1) накапливание искажений, (2) ложные корреляции и (3) побочная эндогенность. Эти вопросы должны заботить тех, кто занимается анализом больших данных даже в том случае, если данные можно считать достоверными. Ошибки, не связанные с выборкой (nonresponse error), только усугубляют эти проблемы.

Чтобы продемонстрировать накапливание искажений (1), представим, что аналитику нужно классифицировать индивидов по двум категориям – K1 и K2 – основываясь на значении 1000 признаков (или переменных) в наборе больших данных. Представим дальше, что исследователю неизвестно, что значения для людей из K1 равны нулю во всех переменных, в то время, как люди из K2 имеют значение, равное трём, только в первых 10 переменных, и значение, равное нулю, в остальных 990. Классификационное правило, основанное на первых переменных, когда $m \leq 10$, работает вполне хорошо и имеет небольшую погрешность. Тем не менее, когда все больше и больше переменных включаются в правило, погрешность растет, поскольку значения не содержат новой информации (то есть эти 990 значений не имеют дискриминационной силы), в конечном итоге перебивая информативные сигналы (то есть первые 10 переменных). В примере Фань с коллегами, когда $m > 200$, накопленные искажения превышают сигнал, заложенный первыми 10 переменными, а само правило классификации становится эквивалентным подбрасыванию монетки в воздух.

Многомерность данных также может приводить к ложным корреляциям (2), когда многие несвязанные признаки оказываются сильно связанными просто случайно, и в результате появляются ложные открытия и ошибочные заключения. Например, используя смоделированные генеральные совокупности и относительно небольшие выборки, Фань и его коллеги [Fan et al., 2014] показали, что при 800 независимых признаках, вероятность аналитика найти точные корреляции с коэффициентом выше 0,4 составляет 50%. Полученные результаты означают, что риски прийти к ошибочным выводам высоки, что связано исключительно с эмпирическим подходом к прогнозной аналитике, основанной на многомерных данных.

И наконец (3) основное ограничение регрессионного анализа заключается в том, что смоделированная независимая переменная не коррелирует с остаточными ошибками. Эндогенность касается как раз нарушения этой аксиомы. Для многомерных моделей такое может произойти случайно — феномен, который Фань и Ляо [Fan, Liao, 2012] называли «побочная эндогенность». Побочная эндогенность приводит к моделированию ложных отклонений в выходных переменных, и, как результат, - к ошибкам в отборе и смещениям в прогнозировании, основанных на модели. Риски побочной эндогенности возрастают с увеличением количества переменных в процессе выбора модели. Соответственно, этот момент особенно важен для аналитиков, работающих с большими данными.

Фан с коллегами [Fan et al., 2014], как и другие авторы (например: [Stock, Watson, 2002; Fan et al., 2009; Hall, Miller, 2009; Fan, Liao, 2012], предлагают продуманные статистические способы, направленные на снижение рисков (1) и (3). Тем не менее, как уже было отмечено ранее, эти и другие проблемы решатся, когда в данные будет введены ошибки, не связанные с выборкой. Бимер и Тревин [Biemer, Trewin, 1991] показали, что такие ошибки вызовут смещения в результатах традиционного анализа данных и увеличат разброс оценочных показателей таким образом, что их будет сложно оценить или уменьшить в процессе анализа. Так, массивность и многомерность больших данных, а также риски случайных ошибок и ошибок, не связанных с выборкой, требуют новых, продуманных подходов к анализу данных.

5.1. Концепция общей ошибки для больших данных

Лучшее понимание источников и природы ошибок, не связанных с выборкой, может упростить работу с рисками, которые возникают в больших данных. Чтобы получить такого рода информацию, необходимо глубокое понимание механизмов генерации данных, инфраструктуры их обработки и подходов, которые были использованы для создания особого массива данных или расчётов, полученных из этих массивов. В данных опроса эта информация закреплена в рамке «общей ошибки опроса», которая распознает все возможные источники ошибок данных, влияющие на валидность и правильность оценки (см., например, [Biemer, 2010]). Концепция общей ошибки опроса помогает объяснить характер источников ошибки, а также то, каким образом эти ошибки влияют на выводы. Концепция раскладывает общую ошибку на смещения и отклонения, которые в дальнейшем можно разбить на элементы, что позволяет рассматривать особые типы ошибок. Важно отметить, что несмотря на то, что наше рассуждение на тему вопросов, связанных с ошибками, подразумевает количественный анализ, некоторые из вопросов, которые здесь рассматриваются, также относятся и к сфере качественного анализа больших данных.

Что касается опросов, концепция общей ошибки опроса даёт полезную информацию о том, какие из шагов генерации данных и подготовительных операций влияют на итоговые расчеты и выводы. Кроме того, такая концепция может подтолкнуть к формированию методов, направленных на уменьшение ошибок, либо помочь сформировать варианты того, как можно скорректировать эффекты этих ошибок в конечных данных, что сделает логические выводы более качественными. Мы считаем, что большим данным необходима концепция общих ошибок. В этом разделе мы предлагаем набросок такой концепции. Наши предложения смоделированы по образцу теории общей ошибки опроса, поскольку, на наш взгляд, многие источники ошибок для обоих случаев одинаковы. Тем не менее, концепция общей ошибки больших данных (Big Data Total Error, BDTE) обязательно должна включать дополнительные источники ошибок, которые присущи только большим данным и могут вызвать существенные смещения и погрешности в результатах. Концепция общей ошибки больших данных, как и теория общей ошибки опроса, будет способствовать нашему пониманию ограничений в данных, что приведет к более качественному анализу и корректной интерпретации результатов. Возможно, это также наполнит содержанием текущие исследовательские задачи по снижению эффектов ошибок в анализе больших данных.

На рисунке 5 приведен типичный массив данных опроса — матрицы, состоящей из некоторого числа строк и столбцов. Массивы данных, полученные из больших данных, могут быть представлены таким же образом, и, соответственно, будут обладать многими схожими свойствами. В соцопросах строки могут быть элементами выборки или всего населения, столбцы — характеристиками элементов строки, ячейки — содержать значения характеристик для каждого элемента. Общая ошибка для такого массива данных может быть представлена следующей приблизительной формулой:

Общая ошибка = Ошибка строки + Ошибка столбца + Ошибка ячейки.

Record #	V1	V2		Vк
1				
2				
N				

Рисунок 5 - Типичный прямоугольный формат
для традиционного анализа данных

Ошибки строки могут быть трех типов:

- Пропуски – некоторые из элементов выборки не представлены в строках.
- Повторы – некоторые из элементов выборки занимают более одной строки.
- Ошибочные коды – некоторые строки содержат элементы или категории, которые не являются частью выборки.

В массивах данных опроса пропуски могут содержать как неотобранные элементы из выборки, так и представителей групп, намеренно исключенных из выборки. В больших данных избирательность является стандартной формой пропусков. Например, массив данных, содержащий людей, которые делали запросы в Гугле за последнюю неделю, автоматически исключает людей, которые не подходят под данный критерий. В отличие от выборки опроса, это форма неслучайной избирательности. Например, люди, у которых нет доступа к интернету, не попадают в массив данных. Исключение этих людей может вызвать смещение, поскольку люди, имеющие доступ к интернету, могут иметь совсем другие демографические характеристики, чем те, у кого такого доступа нет. Эта проблема близка проблеме неполного покрытия при построении выборки, и зависит от населения, которое исследователь пытается измерить.

Мы также склонны предположить, что массивы больших данных, как например массив, содержащий запросы в Гугле за последнюю неделю, может репрезентировать одного и того же человека в разное время. Люди, которые делают много запросов в течение этого периода, будут представлены непропорционально по отношению к тем, кто делал меньше запросов. Другие ошибочные коды могут возникнуть, когда тот, кто производит поиск, является не человеком, а компьютером, например, в случае с веб-скрапингом (web scraping).

Наиболее частая ошибка в столбцах, вызванная неправильным присвоением меток данным, это ошибка метаданных. Например, бизнес-реестр может содержать столбец с меткой «количество сотрудников», определенных как количество человек в компании, которые получают чек по расчёту заработной платы на месяц вперед. Но столбец содержит сотрудников вне зависимости от того, получили ли они чек в прошлом месяце или нет, включая тех людей, который уволились без оплаты. Такого рода ошибки представляются весьма распространенными при анализе больших данных, поскольку процесс создания массива многоуровневый. Например, данные, сгенерированные из такого источника, как индивидуальный Твиттер, могут пройти несколько преобразований, пока попадут в прямоугольный файл, приведенный на Рис.5. Процесс преобразований может быть весьма сложным, например, включать синтаксический анализ фраз, распознавание слов, классификацию их на предмет соответствия теме исследования и на предмет положительных или негативных отзывов об экономической ситуации.

И существуют значительные риски, что итоговые признаки были неправильно определены или проинтерпретированы аналитиком.

Наконец, ошибки в ячейках могут быть трех типов: ошибка контента, ошибка спецификации и отсутствие данных. Ошибка контента происходит, когда значение в ячейке подходит под описание столбца, но является неправильным. Например, значение подходит под определение «количество сотрудников», но указанное число не согласуется с реальным количеством сотрудников в компании. Ошибки контента могут быть следствием ошибок измерения, ошибок обработки данных (например, ввода, кодирования, редактирования и так далее), ошибок вменения значений или вызваны какими-либо другими причинами. Ошибки спецификации по сути те же ошибки столбца, только для ячейки. Например, столбцу было верно присвоено значение и метка; тем не менее, значения для каких-то компаний, даже если они очень точны, не подходят под требуемое определение. Отсутствующие данные, что понятно по названию, это просто пустая ячейка, которая должна иметь значение. Как пишут Кройтер и Пэнг [Kreuter, Peng, 2014], массивы данных, полученные из больших данных, к сожалению, страдают от всех трех типов ошибок в ячейках, особенно от отсутствующих или неполных данных.

5.2. Расширение концепции для больших данных

Традиционная концепция общей ошибки опроса вполне универсальная для применения к практически любому массиву данных, который отвечает формату, приведенному на рис. 5. Тем не менее, в большинстве практических ситуаций она оказывается достаточно ограниченной, поскольку в ней не делается попытки объяснить ошибки в процессах генерации данных. В некоторых случаях такие процессы являются «черным ящиком», и лучший подход — попытаться оценить качество конечного продукта. Для данных опроса рамка общей ошибки опроса создает довольно полное описание процесса генерации ошибок для данных и теории опроса (см., например, [Biemer, 2010]). Кроме того, предпринимаются попытки объяснить эти процессы с точки зрения систем учета населения и административных записей [Wallgren, Wallgren, 2007]. Но в настоящем документе все же мало внимания посвящено перечислению источников ошибок и процессов генерации ошибок для больших данных. Одним из затруднений для этого является тот факт, что процессы, связанные с генерацией больших данных, настолько же разнообразны, как и сами большие данные. Тем не менее, на этом пути можно продвинуться вперед, если представить необходимую последовательность действий. Обычно она такова:

- **Генерация** – данные генерируются из какого-либо источника случайно или целенаправленно.
    - **Извлечение / Трансформация / Загрузка** – данные сводятся вместе в однородной вычислительной среде на трех стадиях:

- *Стадия извлечения* – данные собираются из своих источников, интерпретируются, обосновываются, отбираются и сохраняются.
- *Стадия трансформации* – данные преобразовываются, кодируются, перекодируются, объединяются / разъединяются и/или редактируются.
- *Стадия загрузки* – данные сводятся вместе и отправляются в хранилище данных.

- **Анализ** – данные перерабатываются в информацию через процессы, включающие:
  - *Фильтрацию (отбор) / сокращение* – ненужные признаки и содержание удаляются; некоторые признаки могут быть объединены в новые признаки; элементы данных могут сокращаться или отбираться для того, чтобы на следующих этапах с ними было легче работать.
  - *Расчеты / Анализ / Визуализацию* – данные анализируются и/или подготавливаются к интерпретации и извлечению информации.

На Рис.6 графически показано, как данные перемещаются на этих этапах. Степень серьёзности ошибок, которые возникают в течение этих процессов, зависят от конкретных источников данных и целей, которые поставлены перед аналитиком. Тем не менее, в общем приближении мы можем рассматривать появление ошибок на каждом этапе.

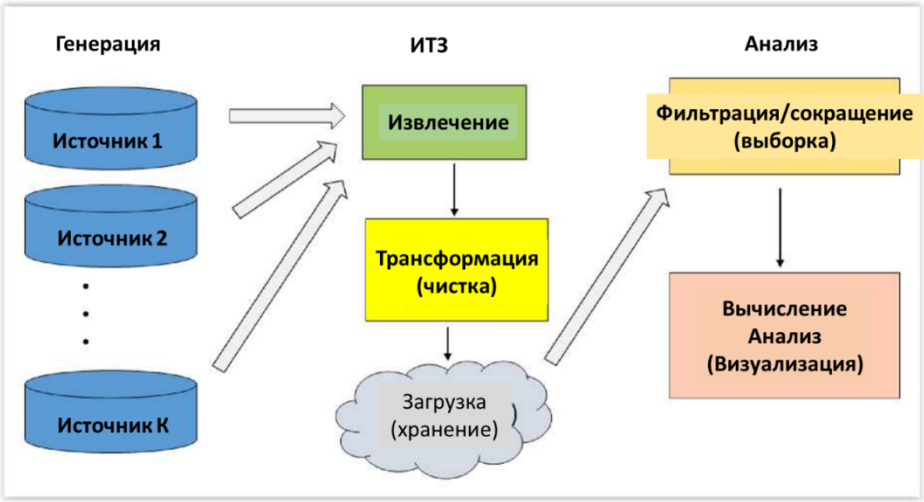

Рисунок 6 - Карта процессов обработки больших данных

Например, ошибка генерации данных очень схожа с ошибками, возникающими при сборе данных опроса. Как и в опросах, универсализация процесса сбора больших данных может вызвать ошибки или пропуски в данных. Кроме того, источники генерации данных могут быть взяты по выбору, и собранные данные не смогут репрезентировать определенные группы людей или тех, кто составляет целевую совокупность. Так, ошибки при генерации данных включают завышение или занижение доли, потерянные знаки, неполные или отсутствующие значения, нарушение случайности отбора, избирательность источников и метаданные, которые не достаточны, отсутствуют или же ошибочны.

Процесс извлечения, преобразования и загрузки данных в чем-то похож на процесс обработки данных на разных стадиях проведения опроса. Последний может включать создание или увеличение количества метаданных, сопоставление с записью, кодирование переменных, редактирование, порчу (или очистку) данных, сведение данных (т.е. установление связей между записями и файлами из разрозненных систем). Ошибки процесса извлечения, преобразования и загрузки данных включают: ошибки спецификации (в том числе ошибки в метаданных), ошибки согласования, кодировки, редактирования, ошибки случайного изменения данных и ошибки сведения данных.

Как уже было отмечено выше, анализ больших данных создает риски накапливания искажений, ложных корреляций и побочной эндогенности, что может быть осложнено ошибками, связанными или не связанными с выборкой. Что касается последних, то чтобы создать репрезентативные или более удобные в работе массивы, можно фильтровать данные, отбирать их или сокращать каким-либо другим способом. Эти процессы могут потребовать дальнейшего преобразования данных. Ошибки включают ошибки выборки, ошибки выборочности (или отсутствия репрезентативности) и ошибки моделирования.

Другие ошибки, которые могут возникнуть на стадии подсчетов, схожи с ошибками оценки и моделирования в опросах. Они включают ошибки моделирования, неадекватные или неправильные корректировки для улучшения репрезентативности, не подходящее или ошибочное взвешивание, ошибки подсчётов и алгоритмов.

В Разделе 4 мы отметили, что все массивы данных имеют риски ошибок, связанные с процессом сбора данных. Американская ассоциация исследователей общественного мнения (AAPOR) призывает к прозрачности этих процессов. Аналогичные усилия были бы очень ценны и в исследованиях, основанных на больших данных.

Выводы

- Использование больших данных статистически обоснованными способами ставит новые задачи. Вера в то, что объем данных может компенсировать любой другой недостаток в данных, ошибочна.
- В общественном восприятии размер часто оказывается важнее вопросов отбора.
- Многие платформы, которые производят статистические расчеты на больших данных, периодически изменяют алгоритмы расчёта (динамические алгоритмы), что приводит к ошибкам в любых долгосрочных исследованиях.
- Крупный размер больших данных может привести к таким проблемам, как накапливание искажений, ложные корреляции и побочная эндогенность.
- На каждом этапе в процессе больших данных генерируются ошибки, которые влияют на оценочные показатели. Каждый источник больших данных имеет свой набор ошибок.
- Необходима общая теория ошибок, когда речь идет об использовании источников больших данных.
- Аналитики больших данных должны осознавать многочисленные ограничения, которые имеют данные, и предпринимать необходимые меры, чтобы ограничить воздействие ошибок больших данных на результаты.
- Многие используемые модели, основанные на больших данных, публично не проверяются на надежность. Потому такие модели остаются весьма ненадежными.
- Инициатива по открытости данных AAPOR может выступить в качестве модели для использования данных вне опросов.

6. Каковы административные, технические и технологические трудности, и как с ними справиться?

Исследование общественного мнения вступает в новую эру, в которой традиционные массовые опросы могут играть менее значимую роль. Распространение новых технологий, таких как мобильные устройства и социальные медиа-платформы меняют социальный ландшафт, который, в свою очередь, является полем деятельности исследователей общественного мнения. По мере того, как скоро распространяются данные технологии, через различные платформы (социальные сети) открывается и доступ к мыслям, чувствам и действиям, воспроизводимым пользователями спонтанно, взаимосвязано и зачастую в публичной форме. Способы, с помощью которых люди получают и обмениваются информацией относительно мнений, взглядов, особенностей поведения, за последнее десятилетие подверглись большей трансформации, чем в любой из предшествующих периодов истории, и эта тенденция, вероятно, будет сохраняться. Широкое распространение социальных медиа и мнений пользователей, высказываемых на этих площадках, подталкивает исследователей к поиску новых инструментов сбора данных, увеличивает дополнительное число источников качественной и количественной информации и в некоторых случаях создает альтернативу традиционным методам сбора данных.

Большие данные предоставляют огромные возможности для разработки инноваций в области исследований общественного мнения. Наряду с традиционными массовыми опросами проведение развернутых обследований благодаря многочисленным ресурсам (таким как социальные сети, мобильные устройства) открывает новые перспективы. Для реализации этих возможностей мы должны систематически решать множество возникающих проблем. В данном разделе рассматриваются некоторые административные проблемы, связанные с использованием больших данных (право собственности, контроль, управление сбором данных, политика конфиденциальности), технические проблемы (многопрофильные профессиональные знания), а также технологические проблемы (вычислительные ресурсы).

6.1. Административная проблема: право собственности на данные

Информация, отражающая ежедневную жизнь людей в цифровом мире (ежедневное использование цифровых технологий), сегодня является полезным ресурсом для исследования. В свою очередь, постоянно возникают затруднения в правовых нормах, вытекающие из отсутствия ясности в том, кому принадлежит информация такого рода — лицу, которое производит информацию, частному лицу или организации, которые собирают эту

информацию (хранители данных), лицу, которое накапливает, анализирует или же преобразует информацию, лицу, заинтересованному в покупке такой информации, или обществу в целом. Проблема осложняется ещё и тем, что эти данные могут рассматриваться одновременно как собственность и как независимые сведения, в зависимости от того или иного закона [Cecil, Eden, 2003]. Новые виды данных подвергают сомнению существующие законы о собственности: данные более не размещаются в статистических агентствах с четко определенными правилами обращения с ними, но генерируются и хранятся на предприятиях или в административных органах. Кроме того, поскольку цифровые данные могут храниться вечно, право на собственность может быть заявлено еще не родившимися родственниками, личная жизнь которых может оказаться под угрозой, в случае раскрытия информации о кровном родстве. Для членов Ассоциации важно оставаться в курсе любых изменений в законах, а также быть осведомленными о различиях в системах законодательств разных стран.

6.2. Административная проблема: Контроль данных

Ярким подтверждением статистической конфиденциальности является формулировка «контроль данных, предназначенных для статистических целей» [Duncan et al., 2011]. Статистические органы занимают ведущее положение в развитии контроля данных по нескольким направлениям. Во-первых, по направлению организации и проведения профессионального обучения сотрудников статистических агентств. Во-вторых, академические программы, (такие как Совместная программа по методологии исследования), организации (такие как Федеральный комитет по статистической методологии) и интеллектуальные ресурсы (Комитет по Национальной статистике) в США в значительной степени поддерживаются федеральным статистическим ведомством. В прошлом внимание сосредотачивалось почти исключительно на разработке методологии для оптимизации аналитической работы с данными обследований, нежели с административными данными. Теперь необходимо внедрять программы по подготовке кадров с целью развития понимания специалистами таких вопросов, как определение целевой группы населения и применение методов кластеризации. Несколько подобных программ появляются и за пределами США. Однако важно интегрировать обучение этим навыкам в уже существующие программы, в частности, если речь идет о тенденции объединения данных, полученных опросными и неопросными методами в ходе одного исследования (см. Раздел 7).

6.3. Административная проблема: Управление сбором данных

Будучи основными учреждениями по сбору данных, статистические организации осуществляли свою деятельность в соответствии с нормами

законодательства. Так, например, Раздел 26 (Федеральная налоговая служба США) и Раздел 13 (Бюро переписи населения) кодекса законов США предусматривают штрафы за нарушение конфиденциальности, в связи с чем агентства разработали условия доступа исследователей в соответствии с установленными законом ограничениями.

Установленные законом допущения не решают проблему использования новых технологий для сбора данных. В частности, четвертая поправка к Конституции ограничивает право властей на «запрос» персональных данных о населении, содержащих информацию о «лицах, движимом и недвижимом имуществе, документах». Противоправные действия государства, вторгающегося в частную сферу, влекут ответственность за «нарушение права на тайну личной жизни». Тем не менее, генерирование больших данных часто происходит в открытой форме, либо посредством заключения соглашения с предприятием, и, следовательно, остается за пределами этих рамок. Один из базовых вопросов относится к определению границы частной жизни, того, что представляет собой необоснованное вторжение в частную жизнь [Strandburg, 2014]. Данные, полученные при взаимодействии со специалистами (юристами, медицинскими работниками), или в результате компьютерных транзакций с клиентами, регулируются в соответствии с законами, предусматривающими запрос на «информированное согласие», и опираются на принципы добросовестного использования данных (Fair Information Practice Principles, FIPP). Несмотря на данное положение FIPP, применимое к «данным», последние, как правило, сводятся к персональной информации и не связаны с проблемами сбора данных в развернутых обследованиях, которые возникают вследствие применения методов определения местоположения и использования интеллектуальных сетей [Strandburg, 2014].

6.4. Административная проблема: Конфиденциальность и повторная идентификация

Риск повторной идентификации индивидов с помощью минимального набора данных интуитивно очевиден. Действительно, одним из способов формально измерить риск повторной идентификации, связанный с конкретным файлом данных, является измерение вероятности того, что эти данные могут быть сопоставлены с базовым файлом [Winkler, 2005]. Если данные включают прямые идентификаторы, такие как имена, номера социального страхования, идентификационные налоговые номера, риск достаточно высок. Однако весьма затруднительным может оказаться и доступ к таким точным идентификационным сведениям, как физические адреса и IP-адреса. Более того, положения Закона о медицинском страховании и обмене идентификационными данными участвующих при этом сторон (Health Insurance Portability and Accountability Act, HIPAA), в соответствии с Постановлением о конфиденциальности от 2003 года, предусматривают

обязательное удаление 18 различных типов идентификаторов, в том числе менее точных, таких как дата рождения, серийные номера транспортных средств, URL-адреса[19] и звукозаписи голосов. Тем не менее, даже, казалось бы, незначительная информация, подтверждающая тот факт, что существует лишь один человек среди целевой группы с таким набором характеристик, позволяет практически полностью идентифицировать личность. Риск повторной идентификации увеличивается в связи с всеобщей доступностью выявленных данных и быстрым прогрессом в технологии кластеризации [Dwork, 2011]. В силу множества переменных, каждый член целевой группы приобретает неповторимые черты. Ввиду того, что большие данные предаются широкой гласности, нет никакой гарантии для индивида, чтобы избежать отбора и идентификации [Karr, Reiter, 2014]. В этой связи, как указывает Ом [Ohm, 2010], человек со знанием почтового индекса индивида, даты рождения и пола может повторно идентифицировать более 80% пользователей Netflix[20], однако ни одни из этих данных не представляют информацию, позволяющую установить личность (Personally Identifiable Information, PII).

6.5. Административная проблема: отсутствие однозначного определения «оправданных мер»

Ограничения, установленные законодательством, в отношении деятельности таких организаций, как Федеральная налоговая служба и Бюро переписи населения США, ясно дают понять, что эти учреждения как производители данных должны принимать «оправданные меры» для защиты данных, хотя такие меры не определены. Доверие явно зависит от мнения людей о неприкосновенности частной жизни, но эти взгляды быстро меняются [Nissenbaum, 2011]. Ниссенбаум также отмечает, что для многих людей становится сложным понимание того, где заканчиваются старые нормы и начинаются новые, поскольку «базовые ограничения на потоки информации от нас и о нас, вероятно, отвечают не столько социальной, этической и политической логике, сколько логике технических возможностей: тому, что позволяет сеть Интернет» [Nissenbaum, 2011, p. 34]. Тем не менее, есть некоторые доказательства того, что люди не требуют охраны частных данных и с удовольствием ими делятся, в том числе, информацией о личной жизни, при условии, что соблюдаются определенные социальные нормы, например, представленные Гербер [Gerber, 2001]. Есть три фактора, которые влияют на эти нормы: акторы (отправители и получатели информации, или поставщики и пользователи); атрибуты (в особенности, виды информации о поставщиках, в том числе то, как они могут быть изменены или объединены); и условия передачи (принципы, обуславливающие информационные потоки).

[19] Uniform Resource Locator (URL) – адрес страницы в сети Интернет — *Прим. перев.*

[20] Netflix – американская компания-поставщик фильмов и сериалов на основе потокового мультимедиа — *Прим. перев.*

Что мы можем почерпнуть из текущих знаний

Кинни описывает различные механизмы взаимодействия между пользователями и конфиденциальными данными [Kinney et al., 2009]. Как отмечают авторы (см. рис. 7), «существуют три основные формы взаимодействия: прямой доступ, доступ на основании раскрытия данных (открытая информация), а также доступ на основании запросов. Прямой доступ создает наименьшие помехи в цепочке взаимодействия пользователей и конфиденциальных данных. Доступ на основании раскрытия данных опирается на порядок разглашения данных, сокрытых в файлах общего пользования. В режиме доступа на основании запроса пользователи не могут напрямую получить личные данные, но могут сделать запрос в электронном или бумажном виде» [Kinney et al., 2009, p. 127]. Подробные обзоры различных подходов представлены в работах Дункан с коллегами [Duncan et al., 2011] и Прада с коллегами (Prada et al., 2011).

Современная литература по общедоступности статистических данных описывает различные способы обеспечения доступа к минимальному набору данных, но весьма поверхностно освещает процедуры и правила их раскрытия.

6.6. Техническая проблема: Навыки, необходимые для интегрирования больших объемов данных в массовых опросах

В зависимости от масштаба данных может возникнуть нехватка навыков и ресурсов, необходимых для работы с большими данными. В частности, эффективное решение большинства проблем, связанных с использованием больших данных, требует наличия специалистов четырех квалификаций:

- Специалист в предметной области (domain expert). Пользователь, аналитик или руководитель с большим опытом в предметной области, знанием возможностей и ограничений в использовании данных.
- Научный сотрудник (researcher). Член команды со знанием методов формального исследования, включая методологию исследования и анализ статистических данных.
- Программист (computer scientist). Технически подкованный работник с образованием в области программирования и обработки данных.
- Системный администратор (sys admin). Сотрудник, ответственный за определение и сопровождение программного обеспечения, позволяющего производить вычисления большого объёма.

Однако, исходя из нашего опыта, многие компании пытаются обойтись лишь одним сотрудником.

Рисунок 7 - Модели взаимодействия данных пользователя [Kinney et al., 2009].

Опыт в предметной области играет важную роль в работе с новыми типами данных, собранных без помощи контрольно-измерительных средств, как правило, в рамках качественных опросов. Например, с точки зрения социальных сетевых сервисов использование больших данных требует глубокого понимания технических возможностей и особенностей поведения пользователей той или иной социальной сети. Так, например, размещение сообщений в Твиттере подчиняется нормам и правилам, которые могут оказывать влияние на интерпретацию данных, полученных из этого источника. Это относится к использованию никнеймов[21] и хэштегов[22], определенных сокращений и терминологии, таких методик, как ретвитинг[23], изменению сообщений (твитов) и возможности ставить отметки «мне нравится». Кроме того, важно понимать, в какой мере различные формы новых медиа не позволяют получить определенные демографические данные (например, малое количество граждан в возрасте 60 лет и старше, использующих Твиттер).

Навыки фундаментального исследования, такие как применение традиционной методологии опросов и использование методов описательной статистики имеют первостепенную важность для понимания больших данных. По мере увеличения объема цифровых данных и уменьшения препятствий в доступе к таким данным, растет и риск низкой подготовки инженеров и программистов, обнаруживающих фиктивные связи в больших данных. Специалистам с традиционной подготовкой необходимо принимать участие в исследовательском процессе на протяжении всех его этапов для того, чтобы должным образом интегрировать большие данные в массовые опросы.

Рисунок 8 - Квалификации, необходимые в работе с большими данными

[21] Никнейм (handle) – имя-псевдоним в социальной сети. — *Прим. перев.*

[22] Хэштег (hashtag) – тематическая ссылка в социальных сетях. — *Прим. перев.*

[23] Ретвитинг (retweeting) – цитирование, перепубликация сообщений в ленте Твиттер. — *Прим. перев.*

Компетенции в области вычислительной техники включают в себя умения работать в командной строке, применять языки программирования, использовать базы данных и языки баз данных, а также опыт работы с передовыми аналитическими инструментами. Чем больше увеличивается набор данных, тем большее значение приобретают навыки работы с базами и знание аналитических методов. Некоторые исследователи предпочитают сотрудничать с квалифицированными программистами, чтобы приобрести необходимый практический опыт. По мере того, как такое взаимодействие приводит к созданию прочных партнерских отношений в исследовательской деятельности, возникает необходимость в организации совместной работы междисциплинарного характера. Основные компоненты информационных технологий, которые часто используются в процессе сбора, хранения и анализа больших данных, включают в себя:

- **Apache Hadoop.** Программное обеспечение для разработки и использования распределённой файловой системы, позволяющей хранить данные больших объемов (терабайты или петабайты содержимого), и параллельной обработки алгоритмов в отношении больших данных. В качестве языков программирования используются Java или Python.
- **Apache Spark.** Быстрый и универсальный инструмент обработки больших объемов данных, обеспечивающий поддержку Hadoop или баз данных в оперативной памяти. Поддерживает языки программирования Java или Python.
- **Язык программирования Java.** Универсальный инженерный язык, обеспечивающий создание эффективных алгоритмов для анализа данных.
- **Язык программирования Python.** Универсальный инженерный язык, обеспечивающий проектирование и создание прототипов, эффективные алгоритмы для анализа данных.

Необходимо заметить, что существует множество различных программ. Несмотря на то, что сегодня широко используется такой инструмент как Hadoop, учитывая быстрое развитие в этой области, в ближайшее время на смену ему может прийти нечто другое. В связи с этим кажется целесообразным обсуждать вопросы кластеризации и параллельной обработки неструктурированных данных.

Системные администраторы играют важную роль в определении, создании и поддержке основного программного обеспечения для хранения и анализа больших данных. Работа с большими данными часто требует дополнительных вычислительных ресурсов. В зависимости от размера данных, такие ресурсы могут варьироваться от аппаратных и серверных

стеков[24], которые могут управляться не специалистами по информационным технологиям, до очень крупного вычислительного окружения, включающего аппаратные и серверные стеки повышенной мощности, работа с которыми зачастую требует специальной подготовки в области ИТ. В качестве примера, многие университеты организуют вычислительные центры высокой производительности (High Performance Computing Center, HPC), предусматривающие сетевые серверы, программное обеспечение (как, например, Hadoop), а также базы данных и пакет аналитических программ. Системные администраторы, ответственные за обеспечение вычислительных платформ больших данных, зачастую прибегают к следующим аппаратным ресурсам:

- **Вычислительный кластер внутреннего типа (Internal Compute Cluster).** Для длительного хранения уникальных или конфиденциальных данных необходимо создавать и администрировать кластер Apache Hadoop с помощью ряда сетевых серверов, организованных в самостоятельную сеть. Использование данного аппаратного ресурса является убыточным с точки зрения краткосрочной перспективы, но выгодным с точки зрения долгосрочной.
- **Вычислительный кластер внешнего типа (External Compute Cluster).** В сфере информационных технологий развивается тенденция применения информационных серверов, предоставляемых в качестве «коммунальных услуг» внешними провайдерами. Такие организации, как Amazon Web Services (AWS) – подразделение Amazon.com – позволяет системным администраторам абонировать предварительно подготовленные кластеры Apache Hadoop, а также системы хранения данных. Параметры установки данного ресурса очень простые, однако его использование может обходиться намного дороже, чем создание долгосрочного кластера внутреннего типа. Среди информационных серверов, аналогичных по функциям *Amazon Elastic MapReduce* выделяются Microsoft HDInsight и Rackspace's Cloud Big Data Platform. Другие серверы системы Hadoop поддерживаются облачными платформами Google и Qubole.
- **Вычислительный кластер смешанного типа (Hybrid Compute Cluster).** В рамках данной опции видится возможным использование вычислительного кластера внешнего вида (как, например, AWS) в целях анализа больших данных, а также создания внутреннего кластера малого размера для длительного хранения данных.

[24] Стек (stack) - структура данных с методом доступа к элементам LIFO (англ. Last In — First Out, «последним пришёл — первым вышел») — *Прим. перев.*

6.7. Технологическая проблема: Требования к вычислениям

Формула измерения расстояния «расстояние = скорость х время» хорошо знакома по урокам математики в средней школе. Данная формула может стать ключом для упрощения понимания того, почему крупные кластеры параллельной обработки вычислительных систем требуются для анализа больших данных. При анализе очень большого набора данных, объем обрабатываемых данных может рассматриваться как расстояние (например, 10 терабайт). Схожим образом, за величину скорости можно принять совокупность процессоров, находящихся в эксплуатации, и жесткие диски – информационные носители.

При прочих равных условиях, система, включающая десять процессоров и десять жестких дисков (10 вычислительных блоков) сможет обработать пакет данных в десять раз быстрее, чем система с одним процессором и одним жестким диском (1 вычислительный блок). Если предположить, что некоторый набор данных состоит из 50 миллионов записей, а системы с одним вычислительным блоком могут обрабатывать 100 записей в секунду, понадобится приблизительно 5 дней и 17 часов (5 000 0000 записей / 100 записей в секунду), чтобы завершить анализ данных – теоретически неприемлемое время ожидания. Система, включающая 10 вычислительных блоков, может произвести обработку с тем же результатом всего за 13 часов 54 минуты, что позволяет значительно сэкономить время. Такие системы, как Apache Hadoop существенно упрощают процесс подключения нескольких компьютеров в кластер, способный поддерживать такие параллельные вычисления.

Несмотря на то, что дисковое пространство обходится относительно недорого, стоимость создания и последующего содержания систем для анализа больших данных может быть довольно высокой. За последние тридцать лет стоимость хранения данных на магнитных носителях, таких как жесткие диски, резко снизилась. Сегодня цена жесткого диска емкостью 3 Тб не превышает 100$ в Соединенных Штатах. Однако общая стоимость системы анализа больших данных складывается из расходов на, как минимум:

- жесткий диск, обеспечивающий хранение информации;
- активные вычислительные компоненты (центральный процессор (CPU)[25], оперативное запоминающее устройство (RAM)[26];
- элементы инфраструктуры (серверная ферма, электричество, система охлаждения, сетевой доступ и безопасность).

[25] Computer Processing Unit (CPU) – центральный процессор

[26] Random Access Memory (RAM) – оперативное запоминающее устройство

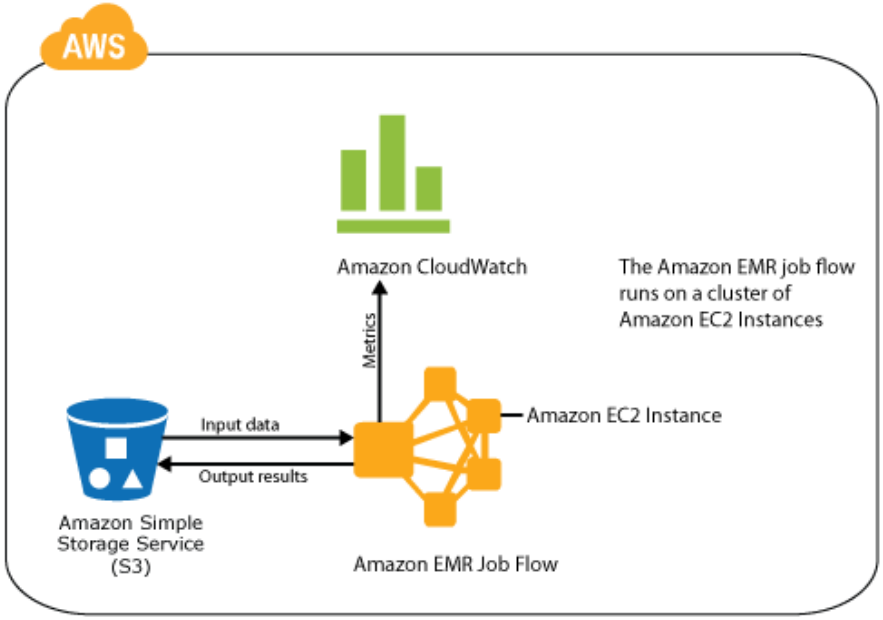

Рисунок 9 - Сервис Amazon Elastic MapReduce (EMR) остается одной из самых популярных утилит облачной платформы Hadoop

В совокупности данные компоненты могут стоить десятки, а то и сотни тысяч долларов. Поэтому экономически не целесообразно создавать компьютерный кластер для работы с большими данными в рамках единичного исследования. Деятельность AAPOR предусматривает возможность формирования взаимовыгодных отношений между государственным и частным секторами не только для обмена данными, но и для совместного пользования аналитической инфраструктурой.

Выводы

- Право собственности на данные в XXI веке не имеет чёткого определения и единого стандарта. Исследователям необходимо тщательно рассмотреть вопросы собственности на данные любого содержания для последующего анализа.
- Мы должны обратиться к дополнительным источникам знаний о том, как обеспечить и хранение больших данных.
- Не существует нормативно-правовой базы для сбора и дальнейшего использования больших данных. Большинство потребителей цифровых услуг (пользователи смартфонов) имеют смутные представления или вообще не догадываются, что данные о них и их действиях могут быть повторно использованы для других целей.

- Удаление ключевых переменных, в том числе информации, позволяющей установить личность (PII), не является достаточным условием для защиты данных от повторной идентификации. Метаданные, содержащие информацию о местонахождении и времени, в совокупности с другими факторами во многих случаях позволяют повторно идентифицировать человека из ранее «обезличенной» записи. Возникает необходимость в разработке новых моделей защиты конфиденциальности.
- Современная литература по общедоступности статистических данных описывает различные способы обеспечения доступа к минимальному набору данных, но менее подробно освещает указания по их раскрытию.
- Эффективное использование больших данных возможно благодаря многопрофильной исследовательской команде, включающей специалиста в предметной области, научного сотрудника, программиста и системного администратора. Многие компании, однако, пытаются обойтись лишь одним сотрудником.
- Организации, нацеленные на пробное использование технологии компьютерной кластеризации с большими данными, могут существенно сократить свои первоначальные затраты, приобретая готовый ресурс во временное пользование (например, Apache Hadoop) у сетевых провайдеров, таких как веб-служба Amazon.
- Такие системы как Apache Hadoop серьезно упрощают создание вычислительных кластеров, способных поддерживать параллельную вычислительную обработку больших данных.
- Несмотря на то, что стоимость жестких дисков снижается, затраты на создание системы для длительного хранения и анализа больших данных остаются высокими. Использование вычислительных кластеров внешнего типа является краткосрочным решением этой проблемы.

7. Как большие данные могут помочь понять происходящее?

Современная литература о развитии больших данных, вероятно, создает впечатление бесконечного соревнования равных по силе традиционных методов исследования, опирающихся на специально предназначенные для научных исследований данные, и новых методов исследования, основанных на «органических» или естественных данных. Исследователи, построившие свою карьеру на анализе данных массовых опросов, особенно обеспокоены ростом больших данных, опасаясь того, что методики, разрабатываемые ими на протяжении всей карьеры, будут устаревать по мере того, как большие данные начнут замещать данные опросов в дальнейших исследованиях.

Аналогичные споры возникают и в отношении статистических методов. Основополагающей теорией для исследований общественного мнения является теория Неймана-Пирсона. Данная теория утверждает, что выборка в исследовании генерируется посредством многократного вероятностного процесса и регулируется зафиксированными в рамках этого процесса базовыми параметрами. Эта концепция также называется частотным подходом к вероятностям, и она наиболее известна большинству исследователей. В качестве альтернативной теории выступает теория байесовской вероятности, предложенная Байесом, Сэведжем, де Финетти и др. В этой теории данные реализованной выборки рассматриваются как фиксированные, в то время как параметры выборки неизвестны и описываются вероятностным образом. Зачастую априорное распределение параметра объединяется с данными наблюдений, обуславливая тем самым апостериорное распределение. Многочисленные обсуждения этих теорий привели к преодолению разногласий и переходу к более практическим позициям. Задача специалистов по статистике — сделать наиболее обоснованные выводы о конечной совокупности, и тем самым создать перспективы для обеих точек зрения. Как частотная, так и байесовская статистика играют ключевую роль в анализе больших данных. Например, в тех случаях, когда наборы данных настолько велики, что анализ выполняется при помощи комплекса вычислительных средств, байесовская статистика предлагает эффективные алгоритмы для объединения результатов такого анализа (см., например: [Ibrahim, Chen 2000; Scott et al. 2013]). Методы организации выборки являются ключевыми в сборе больших данных и для анализа больших данных в рамках малого вычислительного окружения [Leek 2014a; Leek 2014b].

В целом, рассматривать конкуренцию между большими данными и данными опросов или традиционными исследовательскими методами видится нецелесообразным. Необходимо признать совершенствование методологии исследований за счет использования всех форм данных, в том числе больших данных, а также данных, спроектированных с учётом возможных исследований. Безусловно, доступность различных форм больших

данных, рассмотренных в разделе 3, приведет к постепенному уменьшению роли данных опросов. Тем не менее, как большие данные, так и данные опроса имеют свои преимущества и недостатки, которые мы рассмотрим ниже. Эффективной и действенной исследовательской стратегией является рефлексия всех их свойств, проявляющихся в различных условиях, а также внедрение смешанных методов исследования, позволяющих привести точные сведения по интересующим вопросам относительно надлежащего распределения ресурсов.

Любое исследование состоит из вопросов, и один из способов ответа на них – обработка всей доступной информации. Доступность больших данных, которые могут подтвердить результаты исследования, позволяет не только по-новому подойти к старым вопросам, но и заняться новыми, которые до этого были вне поле зрения. Однако решения, основанные на больших данных, неминуемо порождают много вопросов, некоторые из которых лучше решаются традиционными опросными методами. Наряду с широким распространением больших данных параллельно будет увеличиваться и потребность в массовых опросах для решения проблем, поднятых результатами использования больших данных. Доступ к большим данным открывает новые возможности для исследований, в частности, исследователям больше не нужно будет проводить новый опрос в рамках каждого исследовательского проекта. Большие данные могут применяться в целях обеспечения постоянного потока информации о происходящем, например, о поведении потребителей, в то время как методы традиционных исследований позволяют отвечать на более сложные вопросы о том, почему мы наблюдаем определенные тенденции или отклонения от них, например, почему потребители ведут себя так, а не иначе, и что можно сделать, чтобы изменить их поведение.

В размышлениях о том, как комбинировать большие данные с методами традиционного исследования, важно иметь четкое представление о соответствующих проблемах, которые надо будет решить. Большие данные могут быть особенно полезны для выявления статистических закономерностей в данных или построения корреляции между факторами. Напротив, установление причинно-следственной связи между переменными требует того, чтобы данные были собраны в соответствии с определенными моделями и схемами исследований, предназначенными для выделения каузальности. Исследователи маркетинга используют большие данные для так называемого A/B тестирования[27], чтобы установить причинно-следственную связь, что может вызвать определенные сложности в силу использования cookie-файлов. В государственном секторе традиционные исследования, основанные на сконструированных данных, вероятно, сохранят свое первостепенное значение для разработки и оценки политики, в частности,

[27] A/B тестирование (testing) - метод веб-оптимизации; используется для того, чтобы узнать, какой вариант веб-страницы лучше всего конвертирует трафик в действия — *Прим. перев.*

когда на основе специализированных типов данных и схем исследований выстраивается зависимость между причинами общественных изменений и их последствиями. Кроме того, большие данные отвечают потребностям администраторов программ, в чьи задачи входит контроль, обслуживание и улучшение работы программы в рамках существующего политического режима. В этих условиях оценка тенденций, построение корреляций и составление прогнозов зачастую могут оказаться достаточными, поскольку отсутствует необходимость в каузальности, и в таких случаях административные данные и соответствующие источники больших данных подходят наилучшим образом. Однако, когда каузальность не учитывается, а внимание фокусируется на прогнозах, в построении которых используются исключительно накопленные данные, велик риск переноса информации из прошлого в будущее, тем самым встраивая расизм, сексизм и другие спорные направления в современные модели.

7.1. Относительные преимущества данных массовых опросов и больших данных в содействии научным исследованиям

На протяжении многих лет исследования продолжают опираться на данные, собранные в ходе опросов, за неимением других ресурсов. Даже теперь, с развитием альтернативных источников, данные опросов сохраняют значительные преимущества в социальных исследованиях. Основным преимуществом является контроль данных, доступный для исследователей — опрос может быть разработан исключительно в интересах научного исследования. Массовые опросы позволяют кастомизировать показатели результатов по наиболее важным вопросам, которые предстоит решить в ходе исследования. Например, если исследовательский проект направлен, главным образом, на определение размера почасовой заработной платы, целесообразнее провести вспомогательный опрос для измерения этого показателя, а не использовать прокси-сервер или приписывать данные из уже существующих источников.

Контроль, обусловленный массовыми опросами для поддержки научных исследований, также позволяет генерировать оценки, полученные на основе выборки, репрезентирующей интересующую группу населения. При опросе конкретной группы населения в целях создания вероятностной выборки исследователи могут использовать данные об изучаемой группе, чтобы получить оценки, применимые к населению с известной степенью точности. Исследователями подробно разработана теория и практика использования вероятностной выборки и статистического вывода для работы только с этим типом сбора данных, а также эффективного применения этих данных для решения интересующих вопросов.

В отличие от этого, исследователи, опирающиеся на большие данные, получают сведения о населении, представленные этими данными, так как

большинство источников больших данных являются органическими и не зависят от исследователей. Во многих случаях сведения о населении, представленные большими данными, не точно соответствуют интересующей группе – объекту исследования. Например, базы данных, основанные на поисковой системе Гугл, ограничены результатами поиска, проведенного пользователями Гугл, и не отвечают запросам исследователей по населению в целом или его определенной группе. Сложно оценить, в какой степени это может влиять на смещения в результатах относительно того или иного исследовательского вопроса. Исследование телевизионной аудитории и зрительских привычек в Великобритании предлагает выбор между исследованием с выборкой 5100 домохозяйств, репрезентирующей население Великобритании, составленной при поддержке Бюро по исследованию аудитории средств вещания (BARB), и панельным исследованием, проводимом при поддержке телевизионного провайдера SkyView с выборкой 33000 домохозяйств, разработанном SKY Media с помощью опции Sky Digital Homes (выборка домов, подключенных к этой услуге). И хотя панель SkyView значительно превышает размеры панели BARB, последняя позволяет получить оценки, непосредственно репрезентирующие британское население. Другой пример заключается в оценке телевизионной аудитории по каналам Твиттера. Сложность состоит в том, что люди не публикуют ссылку на новостной канал, поведавший им какую-нибудь новость – только сами новости; в свою очередь, зрители отправляют твиты о сериале, который они смотрят в настоящий момент (как, например, «Карточный домик»). В этом случае рейтинги телевизионных новостей будут занижены показателями Твиттера.

Несмотря на это, большие данные имеют ряд преимуществ по сравнению с данными, полученными в массовых опросах, главным из которых является то, что эти данные уже существуют в той или иной форме. Таким образом, исследования, основанные на больших данных, не требуют сбора новых первичных данных. Сбор первичных данных порой оказывается дорогим и медленным, задерживая тем самым обновленные результаты или не позволяя провести новое исследование в силу его высокой стоимости. Проблемы могут также возникать из-за снижения уровня ответов, особенно в условиях проведения длительных обследований.

Сбор и подготовка к анализу больших данных, как правило, требуют меньше усилий и времени, по сравнению с данными массовых опросов. Однако создание комплекса больших данных для дальнейшего анализа – далеко не элементарный процесс. Несмотря на уже имеющиеся большие данные, процесс сбора и объединения цифровых данных из различных источников предполагает сложную исследовательскую работу. По оценкам экспертов, специалисты по обработке данных уделяют 50-80 процентов времени сбору и верификации для последующего анализа [Lohr, 2014]. Задачи анализа больших объемов данных зачастую включают в себя сбор данных из различных источников, и эти наборы данных – как считанных с

датчиков, взятых из документов, найденных в сети, так и полученных более традиционными методами – поступают в различных форматах. Как следствие, новые компании занимаются разработкой программного обеспечения для автоматизации сбора, удаления, хранения и сортировки данных из различных источников, с целью освобождения специалистов от выполнения однообразных заданий, связанных с подготовкой данных. Однако, поскольку каждое исследование предусматривает определенный тип обработки данных, необходимость в такой рутинной работе не исчезнет.

Большие данные зачастую представлены в больших объемах, и с учетом современных технологий, эти большие объемы данных сегодня легче обрабатывать, хранить и изучать, чем в прошлом. В течение многих лет исследователи работали с наборами данных, полученных в результате сотен или тысяч обследований, образующих относительно простую прямоугольную структуру, с n наблюдениями и k переменными. Наряду с тем, что эти наборы данных характеризуются простотой в использовании, ограниченный объем создает препятствия в охвате и широте статистического анализа. В отличие от них большие данные поступают в различных форматах, а возможности больших объемов наблюдений не заставляют заботиться о статистической значимости, как то было в прежние времена. Однако, как уже упоминалось в Разделе 5, большие объемы имеют свои недостатки. Неупорядоченная структура (или ее отсутствие) и огромные объемы больших данных могут оказаться проблемой для обработки и организации информации, в то же время их возможности позволяют получить более полное и детальное представление о процессах. Наиболее исчерпывающие данные могут стать предпосылками для новых вопросов и исследовательских проектов, которые будут информировать нас о различных мероприятиях и последствиях экономических изменений. Наконец, высокая детализация позволяет исследователям подробно изучить особенности поведения, а также подгруппы населения с достаточной статистической значимостью. Например, с помощью традиционных методов исследования можно измерить влияние размера класса на успеваемость учащихся, однако большие данные позволяют нам исследовать данный вопрос в корреляционной зависимости от этапа обучения, школы, учителя или состава класса; предполагается, что все факторы, искажающие результаты, не будут приниматься в расчет. Кроме того, в Разделе 4 настоящего доклада также говорится об использовании больших данных для изучения областей больших отклонений, которое видится невозможным с небольшим набором данных.

Большие данные обычно доступны в режиме реального времени, так как выстраиваются органически, наряду с последовательными действиями индивида (например, телефонные звонки, поиск и покупка в сети Интернет, и т.д.). За счет такой особенности, большие данные приобрели популярность в частном секторе, где компании могут использовать данные для принятия управленческих решений в установленные сроки. Традиционное исследование, опирающееся на сбор первичных данных, осуществляется

медленно, и, соответственно, не может обеспечить оперативное принятие решений. Как выразился один из аналитиков, традиционное исследование разработано для «удобства, а не скорости», оно генерирует качественные результаты, тем самым укрепляя уверенность в исследовании, но генерирует их слишком медленно, поэтому не обеспечивает своевременное принятие решений. Напротив, сроки обработки больших данных более соответствуют темпу принятия управленческих решений в частном или государственном секторе – там, где предусмотрена система премиальных надбавок за эффективную выработку решений, отвечающих быстрым изменениям потребительского спроса или потребностей граждан.

7.2. Методы исследования, опирающиеся на большие данные

Как обсуждалось выше, использование больших данных является наиболее предпочтительным в тех случаях, когда руководство заинтересовано в поиске информации, необходимой для принятия принципиальных решений. Такой организации, нуждающейся в обработке больших данных для обеспечения эффективной работы одной или нескольких программ, можно предложить как минимум три способа решения этой задачи. Во-первых, большие данные могут помочь в подборе людей и подходящих для них программ. Например, работодатель, участвующий в программе по охране труда в целях улучшения здоровья сотрудников, хотел бы отправлять их именно в те учреждения, которые могут оказать требующуюся им помощь, для чего потребуется сбор, обработка и анализ данных об индивидуальных показателях здоровья и поведении этих сотрудников. Во-вторых, большие данные могут быть использованы в целях упрощения взаимных расчетов. В случае с программой по охране труда это могло бы быть равнозначно использованию больших данных для формирования и укрепления здорового профессионального и личного взаимодействия между сотрудниками, медицинскими работниками и страховщиками. В-третьих, большие данные могут быть использованы для определения достигнутых результатов и влияния программы на достижение этих результатов участниками, а также оценки её чистой стоимости. В случае программы по охране труда сотрудников, с помощью больших данных можно измерить состояние здоровья и результаты труда, экстраполируя показатели на будущие результаты, оценить влияние программы на эти результаты и монетизировать оценки воздействия с целью определения чистой стоимости программного инвестирования. На основании этих оценок менеджеры могут принимать взвешенные решения, связанные с развитием программы в долгосрочной перспективе, с тем, чтобы в полной мере обеспечивать потребности работников и работодателя. Конечно, в каждом из этих примеров есть риск того, что использованная информация не будет отвечать интересам работников, что возвращает нас к ранее обсуждаемым этическим проблемам.

Частные компании, учитывая потенциальную пользу больших данных в принятии научно обоснованных решений, в большей степени склонны полагаться на большие данные и методы исследования, опирающиеся на преимущества этих данных. Предиктивная аналитика и методы быстрой цикличной оценки – два метода исследования, основанные на технологии больших данных, и приобретшие популярность в частном секторе в последние годы. Эти методы позволяют менеджерам не только отслеживать текущую деятельность, но и обеспечивать принятие решений относительно того, как тактически реагировать на изменения конъюнктуры и потребительской базы.

Прогнозная аналитика включает широкий спектр методов, используемых для прогнозирования результатов. Например, в частном секторе метод прогнозной аналитики позволяет спрогнозировать реакцию потребителей (в том числе, потенциальных) на определенные изменения, такие как изменение продукта или услуги, внедрение новых маркетинговых мер, открытие нового магазина или появление нового продукта или услуги. Коммерческие компании могут использовать этот метод, чтобы оценить возможное влияние данного изменения на производительность, удовлетворенность клиентов и прибыльность, и тем самым избежать дорогостоящих ошибок. Прогнозная аналитика может проводиться на основе данных, которые собираются в рамках обычной экономической деятельности и хранятся с целью обеспечения текущего анализа. Эти данные могут также объединяться с другими источниками больших данных или данными массовых опросов, полученными извне.

Кроме того, в последние годы моделирование, основанное на прогнозной аналитике, способствует обеспечению новых информационных продуктов и услуг. Например, рекомендации сервисов Amazon и Netflix опираются на прогнозные модели тех или иных фильмов и книг, которые человек хотел бы приобрести. Результаты запросов в Google и новостная лента зависят от алгоритмов, предугадывающих актуальность конкретных веб-страниц или статей. Прогнозная аналитика также используется компаниями в целях профилирования потребителей и подбора соответствующих услуг. Например, страховые компании используют предиктивные модели для «оценки риска» физических лиц на основании их характеристик и истории болезни с целью регулирования платежей. Аналогичным образом прогнозные модели нарушения сроков платежей и погашения применяются кредитными компаниями, что позволяет им контролировать страховую и маркетинговую деятельность, а также ценовую политику.

Методы быстрой цикличной оценки суть копия предиктивной аналитики в ретроспективном ракурсе – используются для быстрой оценки влияния того или иного изменения на результаты производительности, удовлетворенности потребителей и прибыльности. Как и предиктивный метод, быстрая цикличная оценка операционализирует необходимые данные, а также использует другие источники больших данных. Точные статистические

методы, применимые в рамках быстрой цикличной оценки варьируются в зависимости от предпочтений и ресурсов пользователя. Например, быстрая цикличная оценка может опираться на экспериментальные методики, согласно которым изменение происходит в случайно выбранных сегментах деятельности компании или в случайно отобранной потребительской группе. Таким образом, оценка данного изменения может проводиться методом сравнения результатов «экспериментальной группы», подвергшейся изменению, и «контрольной группы», оставшейся без изменений.

Частные организации вкладывают большие средства в реализацию этих возможностей. Так, например, компания Capital One[28] является пионером в области применения быстрой цикличной оценки, осуществляемой на основании данных их операций для выработки бизнес-решений. Компания проводит более 60 000 экспериментов и выполняет анализ по ряду вопросов, связанных с ее деятельностью или предлагаемой продукцией. Многие другие компании также движутся в этом направлении (Manzi 2012). Несмотря на то, что государственный сектор, в отличие от частного, не торопится внедрять большие данные и техники анализа данных, государственные чиновники начинают все выше оценивать значимость этих методов и экспериментов с их использованием в обеспечении управленческих решений и повышении качества государственных программ (Cody and Asher 2014). На широком уровне некоторые правительственные учреждения в целях выработки чёткой аналитической картины собирают доступные данные и анализируют те из них, что связаны с их деятельностью. Например, одна из последних редакционных статей New York Times (от 19 августа 2014) освещает данную тенденцию в Нью-Йорке, сосредоточив внимание на реализации проекта ClaimStat[29], не так давно разработанного ревизором города Скоттом Стрингером. ClaimStat собирает и анализирует данные о судебных исках, ежегодно предъявляемых городу. Выявляя закономерности в выплатах и работе «проблемных» учреждений, представители городской администрации надеются в дальнейшем использовать эти схемы, чтобы снизить количество будущих исков и затрат по ним (New York Times Editorial Board 2014).

[28] Capital One Financial Corporation – банковская холдинговая компания, которая предлагает различные финансовые продукты и услуги (депозиты, кредитование) частным лицам, малому бизнесу и коммерческим клиентам в Соединенных Штатах, Канаде и Великобритании. Вебсайт: https://www.capitalone.com/

[29] Система ClaimStat, разработанная по аналогии с системой CompStat (сокр. COMPlaint STATistics), позволившей обнародовать статистику преступности и способствовавшей усилению борьбы с преступлениями в наиболее опасных районах, представляет собой базу данных Нью-Йорка, содержащих информацию о внесудебных соглашениях и текущих исках. Данная система призвана обеспечить открытость и доступность информации о жалобах и судебных исках горожан против городских организаций, а также сэкономить средства, выделенные для компенсации истцам. Вебсайт: http://comptroller.nyc.gov/reports/claimstat/

Государственный сектор может обращаться к методам предиктивной аналитики для того, чтобы ориентировать услуги для лиц, нуждающихся в материальной помощи, или прогнозировать реакцию индивидов или подгруппы индивидов на такие государственные меры, как создание новой программы или внесение изменений в существующую программу (Cody and Asher 2014). Например, администраторы программы могут применять административные данные и предиктивные методы, чтобы идентифицировать потребителей, подверженных рискам неблагоприятного исхода (безработица, мошенничество, не вызванная необходимостью госпитализация, смертность или рецидив заболеваний). Определяя участников «группы риска», персонал, осуществляющий руководство программой, может проводить целевые мероприятия, чтобы снизить вероятность неблагоприятного исхода или уменьшить отрицательный эффект такого исхода.

С помощью информации, сгенерированной методами предиктивной аналитики, администраторы имеют возможность определить тех, кто получает пользу от предпринимаемых мер, и разработать пути оптимизации таких мер. Как и в частном секторе, предиктивная аналитика апеллирует к операционным данным в рамках ежедневного администрирования программы, и аналитические материалы встраиваются непосредственно в операционные системы данных, что обеспечивает принятие решений в режиме реального времени. Так, предиктивная аналитика в рамках определенной программы оказывается полезной для рядовых сотрудников социальных служб, поскольку ориентирована на выявление случаев несоответствия услуг предъявляемым требованиям и персонализацию услуг для удовлетворения конкретных потребностей отдельных лиц. В некоторых государственных системах страхования от безработицы применение статистических моделей позволяет выявить новых соискателей, вероятно, оставшихся без работы на длительный период, и направить их в службы занятости. При использовании предиктивных моделей по-прежнему важно соблюдать этические и юридические требования, что, к сожалению, не всегда происходит (обсуждение антиконституционных решений см. http://bit.ly/1EpKt2j).

7.3. Совмещение больших данных с результатами опросов

Несмотря на теоретические и практические преимущества анализа больших данных, описанные выше, предпочтительная стратегия — комбинирование новых и традиционных источников данных, поддерживающих исследовательскую, аналитическую и управленческую активность. Особенности комбинации нового и традиционного зависят от запроса в конкретной ситуации. Как представлено во введении, традиционные исследования, опирающиеся на первичные данные, могут отвечать на вопросы, которые нельзя адекватно или легко решить

посредством обращения к большим данным. Во многих случаях это потребует выхода за пределы наблюдаемых тенденций или поведения, легко фиксируемых при помощи больших данных, и перехода к более систематическому рассмотрению причин их появления. Например, представим крупного рекламодателя, имеющего возможность в реальном времени проводить мониторинг складских запасов и объемов продаж. Традиционный дизайн исследования с опросом панелистов о мотивах покупок и местах их совершения поможет лучше ориентироваться на целевые группы покупателей. Альтернативно можно расширить дизайн исследования, включив в него данные о складских перемещениях и объемах продаж как первичной мониторинговой информации, а опросы, в свою очередь, использовать для уточнения причин тенденций, изменений или аномалий, обнаруженных в первичных мониторинговых данных.

Недавно исследователи сформулировали идеи по совмещению больших данных с традиционными опросами в маркетинге. Например, Дуонг и Миллман [Duong, Millman, 2014], опираясь на экспериментальные данные, подчёркивают, что в оценке брендов работа с потоками интернет активности пользователей может успешно совмещаться с опросами, что позволяет лучше оценивать эффективность рекламы. В их эксперименте данные, описывающие взаимодействие пользователей с вебсайтами, комбинировались с традиционными данными онлайн опроса, что позволило представить эффект от различного вида рекламных сообщений. Аналогично Портер и Лазаро [Porter, Lazaro, 2014] описывают серию бизнес кейс-стади для иллюстрации того, как опросные данные могут комбинироваться с другими источниками информации, тем самым позволяя проводить комплексный анализ. В одном из кейсов авторы описывают смешанные стратегии анализа данных для установления сходств и различий у респондентов. Данные о потребительском поведении, снимаемые через активность на вебсайтах и торговые транзакции, комбинируются с опросными данными о восприятии, установках, жизненных событиях и оффлайн активности. Используя модели индивидуального поведения респондента в отношении его потребительских предпочтений (на основе опросных данных) и поведения (данные об интернет активности), они лучше интерпретируют особенности онлайн активности и выделяют области для дальнейших разработок, направленных на удовлетворение потребностей различных групп покупателей.

Смешанные стратегии также широко применяются государственными агентствами. Например, Национальный центр статистики здоровья (National Center of Health Statistics, NCHS) разрабатывает программу связанных записей, направленную на максимизацию научной ценности данных, основанных на опросах[30]. Программа связывает различные опросы Национального центра статистики здоровья не только с административными

[30] http://1.usa.gov/1IlwiLW

записями из центров по оказанию бесплатной и льготной медицинской помощи (Centre for Medicare and Medicaid Services, CMS) и Администрации социального обеспечения (Social Security Administration, SSA), между которыми все данные согласованы, но и с Секретариатом по планированию и оценке (Office of the Assistant Secretary for Planning and Evaluation), что создаёт возможность применения смешанных стратегий анализа. Например, Дэй и Паркер [Day, Parker, 2013] обращаются к данным программы связанных записей для сопоставления самоотчетов о диабете, получаемых в Национальном опросе о здоровье (National Health Interview Survey, NHIS), с записями о диабете, используемыми в медицинских хранилищах данных (Medicare Chronic Condition Summary file). В конечном счёте, связанные файлы данных позволяют исследователям в деталях изучать факторы, влияющие на инвалидизацию, хронические заболевания, профилактику здоровья, заболеваемость и смертность.

Аналогичным образом Бюро переписи США ищет способы обращения к большим данным для улучшения качества опросов и переписей населения: снижения времени сбора данных, улучшения объяснительных моделей и сокращения операциональных издержек на сбор информации [Bostic, 2013]. Например, Бюро переписи планирует начать использовать данные электронных транзакций и административные записи для дополнения или улучшения статистических данных. В сфере строительства ведётся оценка данных продавцов новостроек о стоимости жилых помещений, реализуемых без права выкупа по закладной, что позволит улучшить статистические оценки нового жилья и продаж жилой недвижимости. Также рассматриваются способы обработки онлайн записей, производимых органами местного самоуправления и государственными агентствами. В секторе розничной торговли Бюро переписи оценивает электронные платежи, чтобы заполнить недостающие данные по географическому распределению продаж и величине доходов в зависимости от размеров компании. Во всех северных европейских странах есть система статистических регистров, которые на регулярной основе используются для производства статистической информации. Система состоит из четырех основных блоков (см. рис. 10): население, активность, недвижимость и реестр юридических лиц [Wallgren, Wallgren, 2014].

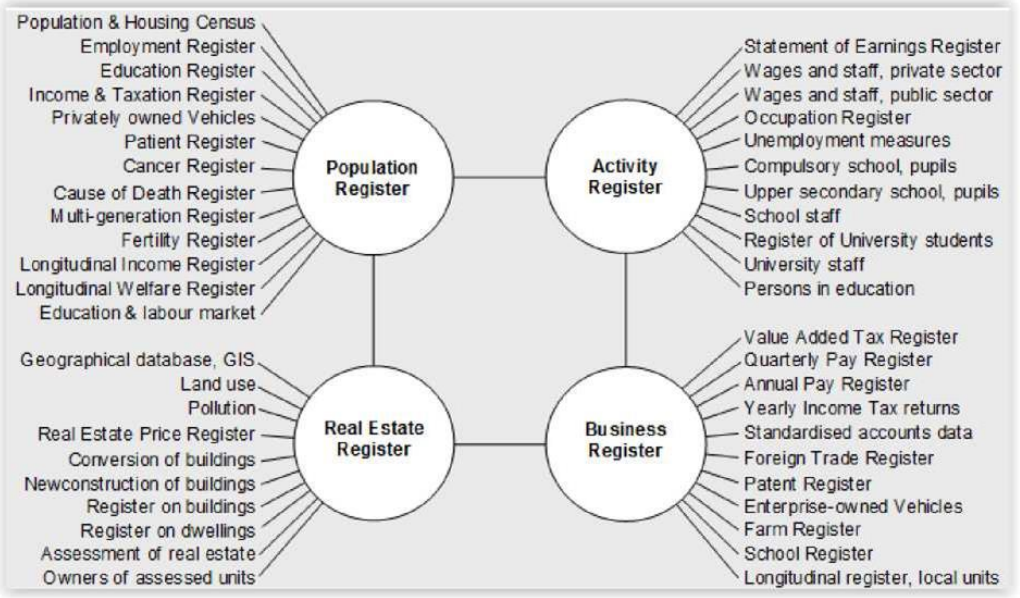

Рисунок 10 - Система статистических регистров по типам объектов и полям субъектов, источник: [Wallgren, Wallgren, 2014]

Выводы

- Опросы и большие данные — комплементарные, а не конкурирующие методы. Существуют большие различия в подходах, но они должны рассматриваться как преимущества, а не как недостатки.
- Исследование — это прежде всего поиск ответов, и лучший способ этого — рассмотрение всех возможных источников информации. Большие данные — один из таких источников, который обеспечивает новые решения прежних вопросов и создаёт новые вопросы.
- В частном секторе большие данные необходимы для управления деятельностью и принятия решений. Примерами подобных исследовательских техник могут выступать предиктивная аналитика и методы быстрой цикличной оценки.
- Аналитические решения для улучшения государственного управления встречаются гораздо реже.
- Большие данные могут быть использованы для обнаружения алгоритмов и выявления корреляций между факторами.

8. Выводы и требования

В этом разделе мы возвращаемся к основным вопросам, поставленным перед проектной группой.

A. Могут/должны ли большие данные использоваться для накопления статистических сведений о знаниях, мнениях и поведении людей?

Существует много разновидностей больших данных (раздел 3). В этот отчёт мы включили административные данные как одну из них. Типы больших данных различаются по объему контроля исследователя над ними и потенциальной значимостью выводов, получаемых от их анализа [Kreuter, Peng, 2014]. На одной стороне континуума у нас административные данные, которые в некоторых странах накапливаются уже долгие годы, например, в северных странах переписи населения основаны на административных записях. Статистические агентства заключают партнёрские соглашения с собственниками административных данных и могут оказывать влияние на дизайн этих данных. На другой стороне континуума — большие данные из платформ социальных медиа, где исследователь не имеет никакого контроля или воздействия на их формирование и накопление. За последние несколько лет уже получены примеры статистических оценок, основанных на данных социальных медиа. Мы также видели исследования, в которых сопоставляются оценки, полученные из больших данных и традиционных опросов. Однако в настоящий момент не накоплено достаточно опыта, который позволил бы сформировать лучшие практики по выделению статистических оценок населения из социальных медиа. Промежуточным решением на континууме могут считаться примеры больших данных, полученных посредством направленного контроля, например, размещения датчиков движения для замеров трафика или дополнительных показателей перемещений.

Одно из наиболее существенных критических замечаний в отношении больших данных связано с отсутствие теории по формированию выводов, получаемых из обработки данных. Хотя многие не перестают повторять, что большие данные, действительно, большие, этого отнюдь не достаточно. Теория выборок, которой сейчас пользуются многие статистические агентства, была сформирована во времена, когда сбор данных проходил исключительно через опрос всего населения. Это было очень дорогостоящее решение, и теория выборок решала проблему стоимости. Сегодня огромное количество больших данных генерируется как сопутствующие продукты, не требующие дополнительных усилий. Одновременно уровень участия в опросах постоянно снижается, стоимость опросов возрастает, и многие положения теории выборки становятся проблематичными из-за роста неответов и других ошибок, не связанных с выборкой. Мы находимся на пути

от традиционной парадигмы опросов к новой парадигме работы с множественными данными. Большие данные — один из возможных источников, который даёт значимую информацию. Важно развивать теорию и методы для полного и комплексного анализа больших данных, особенно того типа, в котором мы не влияем на дизайн сбора данных. Мы ещё в начале пути.

Обнаруженные или собранные большие данные содержат ошибки, которые существенным образом влияют на выводы, и каждый источник больших данных содержит свой набор ошибок. Потенциальные воздействия каждой ошибки будут зависеть от источника больших данных. Как и для «небольших данных», нам следует придерживаться концепции общей ошибки и для больших данных (раздел 5). Концепция общей ошибки больших данных может послужить хорошей рамкой для консолидации усилий исследователей [Biemer, 2014].

B. Как большие данные могут улучшить и/или заменить существующие «классические» методы исследования, такие как опросы и/или переписи населения?

Возможности больших данных определяются как новыми подходами к прежним вопросам, так и постановкой новых вопросов, о которых ранее даже не задумывались (раздел 7). Большие данные могут быть использованы для создания постоянного потока информации о происходящем, например, о потребительском поведении. При этом традиционные исследования могут быть сфокусированы на более детальных вопросах, почему мы наблюдаем те или иные зависимости или отклонения от них, например, почему покупатели ведут себя определенным образом, и как это поведение можно поменять.

Административные данные во многих странах используются в качестве основ выборки, в оценочных процессах — для увеличения точности, в комбинации с опросами — для уменьшения нагрузки на респондентов. Другие виды больших данных могут применяться аналогичным образом. Платформы социальных медиа могут рассматриваться в качестве источников оперативной информации о том, что люди думают о разных концептах, а также в качестве площадок для тестирования вопросов.

Кроме того, административные данные используются в качестве золотого стандарта для некоторых методологических исследований. Например, Дей и Паркер [Day, Parker, 2013] брали данные из программы связанных записей для сопоставления самоописаний диабета в Национальном опросе о здоровье (NHIS) с идентификаторами диабета в медицинских хранилищах данных (Medicare Chronic Condition Summary file).

Если мы выходим за пределы административных записей и смотрим на разные типы больших данных, мы видим противоположную тенденцию. Сейчас опросные данные рассматриваются в качестве стандарта. Во множестве работ, где в качестве источников оценок выступают большие

данные, они сопоставляются с оценками, полученными в традиционных опросах (раздел 3). Корреляции между двумя типами оценок представляют особый интерес в таких работах. Если корреляции высокие (и не подвержены воздействию от изменений неизвестных алгоритмов), статистики больших данных могут использоваться в качестве систем раннего предупреждения (например, гугловский прогноз заболевания гриппом), поскольку они требует меньше временных и материальных затрат. Транспарентность алгоритмов сбора данных — основное условие работы с такими данными. Кроме того, должны быть заключены соглашения с частным сектором для того, чтобы сохранялась уверенность в их стабильности и открытости.

В частном секторе большие данные применяются для управления и принятия решений. Примерами исследовательских приёмов могут считаться предиктивная аналитика и методы быстрой цикличной оценки.

C. Могут ли большие данные опережать опросы? Что делать, если любые текущие обращения к большим данным (изучение публичного знания, мнений или поведения) представляются слишком многообещающими? Какие типы приложений выглядят неадекватными?

Большие данные обладают рядом достоинств по сравнению с данными массовых опросов. Одно из неоспоримых достоинств состоит в том, что большие данные присутствуют до проведения какого-либо исследования и не требуется дополнительных усилий по сбору первичных данных. Сбор первичных данных обычно весьма дорог и требует много времени, что приводит к замедлению в разработке новых исследований или их неоправданному удорожанию.

Как отмечалось выше, административные данные широко используются во многих странах. Северные страны располагают системой статистических регистров, которые регулярно используются для расчета статистических параметров о населении, бизнесе, экономической активности или операций с недвижимостью.

Хорошая стратегия заключается в комбинировании новых и традиционных источников данных, необходимых для исследования, аналитики или принятия решений. Сочетания такой комбинации зависят от потребностей в конкретной ситуации. Сканированные данные розничной торговли — один из примеров больших данных, которые в сочетании с традиционными опросами повышают качество и снижают стоимость исследования. Сканированные данные, например, в некоторых странах используются в расчётах Индекса потребительских цен (Consumer Price Index, CPI). Другой пример – когда большие данные получаются посредством снятия информации из трекингового оборудования. Счетчики шагов из подключенных к сети педометров дают куда более точные данные, нежели традиционные опросы, из-за забывчивости респондентов и невозможности дать точную оценку пройденному расстоянию. Другие источники больших

данных со схожими характеристиками относятся к информации, снимаемой различными сенсорами или получаемой в результате регистрации торговых транзакций. Все эти примеры указывают на то, что комбинация больших и малых данных является наиболее оптимальной стратегией сбора информации. Мы надеемся, что усилия AAPOR и других заинтересованных лиц в этой области расширят возможности интеграции данных.

D. Какие операциональные и статистические проблемы связаны с использованием больших данных?

Текущее развитие больших данных само по себе — проблема. Весьма трудно развивать новую технологию, которая слишком быстро видоизменяется. Хорошая стратегия для изучения больших данных — формировать партнерства, в которых мультидисциплинарные команды смогут оценить и применить весь комплекс достоинств, открываемых с большими данными (раздел 6).

Собственность на данные пока точно не определена. И сейчас нет прозрачных легальных оснований для сбора и последующего использования больших данных. Большинство пользователей электронных устройств не подозревают, что их поведение регистрируется, и получаемая информация может быть использована для других целей. Исследователи должны предельно осторожно подходить к вопросу о собственности данных, какой бы контент они не пытались извлекать для анализа. Удаление ключевых переменных, связанных с персональными данными (Personally Identifiable Information, PII) не столь эффективно для защиты данных от последующей реидентификации. Во многих случаях сочетание геопозиционирования и времени регистрации с другими факторами позволяет провести реидентификацию «анонимизированных» записей. Налицо явная потребность в новой модели защиты частной информации.

Организации, нацеленные на эксперименты с компьютерными технологиями больших данных, могут уменьшить потребности в первоначальном капитале за счет аренды кластерных компьютерных ресурсов (таких как Apache Hadoop) у онлайн провайдеров. Системы, подобные Apache Hadoop, кардинально упрощают создание компьютерных кластеров, позволяющих запускать параллельную обработку больших данных.

Хотя стоимость огромных медиа хранилищ может быть не велика, создание новых систем для долгосрочного хранения и анализа больших данных остается весьма затратной. Использование внешних компьютерных кластеров — одно из краткосрочных решений этой проблемы.

9. Литература

1 Acquisti, A. The economics and behavioral economics of privacy // Privacy, Big Data, and the public good: frameworks for engagement / Ed. by J. Lane, V. Stodden, S. Bender, and H. Nissenbaum. Cambridge: Cambridge University Press, 2014. P. 98-112.

2 Antenucci, D., Cafarella, M., Levenstein, M., Ré, Ch., Shapiro, M.D. Using social media to measure labor market flows / NBER working paper series. 2014. [doi:10.3386/w20010]

3 Barocas, S. Nissenbaum, H. Big Data's end run around anonymity and consent // Privacy, Big Data, and the public good: Frameworks for engagement / Ed. by J. Lane, V. Stodden, S. Bender, and H. Nissenbaum. Cambridge: Cambridge University Press, 2014. P.44-75.

4 Biemer, P.P. Total survey error: Design, implementation, and evaluation // Public Opinion Quarterly. 2010. Vol. 74. No. 5. P. 817-848.

5 Biemer, P.P. Toward a total error framework for Big Data / Presentation at the American Association for Public Opinion Research (AAPOR) 69th Annual Conference, Anaheim, CA. 2014. 17 May.

6 Biemer, P.P., Trewin, D. A review of measurement error effects on the analysis of survey data // Survey measurement and process quality / Ed. by L. Lyberg, P.P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, D. Trewin. New York: Wiley & Sons, 1997. P. 603-632.

7 Bosch ten, O., Windmeijer, D. On the use of internet robots for official statistics / Presentation at the Meeting on the Management of Statistical Information Systems (MSIS). Dublin, Ireland, 2014. 16 April.

8 Bostic, W.G. Big Data Projects at the Census Bureau / Presentation to the Council of Professional Associations on Federal Statistics (COPAFS). Washington, DC, 2013. 1 March.

9 Brynjolfsson, E., Hitt, L.M., Kim, H.H. Strength in numbers: How does data-driven decision-making affect firm performance? / ICIS 2011 Proceedings, paper 13.

10 Butler, D. "When Google got flu wrong // Nature. 2013. No. 494(7436). P. 155-156.

11 Cecil, J., Eden, D. The legal foundations of confidentiality / Cited by Julia Lane. Key Issues in Confidentiality Research: Results of an NSF Workshop. National Science Foundation. 2003. [Online] <http://1.usa.gov/1Eq58Df>. [Retrieved January 28, 2015]

12 Cody, S., Asher, A. Smarter, better, faster: The potential for predictive analytics and rapid-cycle evaluation to improve program development and outcomes / Mathematica Policy Research. 2014. [Online] <http://brook.gs/1AMaW5t> [Retrieved January 28, 2015]

13 Couper, M.P. Is the sky falling? New technology, changing media, and the future of surveys // Survey Research Methods. 2013. Vol. 7. No. 3. P. 145-156.

14 Cukier, K., Mayer-Schoenberger, V. "Rise of Big Data: How it's changing the way we think about the world // Foreign Affairs. 2013. Vol. 92. No. 3.

[Online] <http://fam.ag/1bKTv6t> [Retrieved January 28, 2015]
15 Daas, P.J.H., Puts, M. Social media sentiment and consumer confidence / European Central Bank Statistics Paper Series No. 5.Frankfurt, Germany, 2014.
16 Daas, P.J.H., Puts, M., Buelens, B, van den Hurk, P. Big Data and Official Statistics / Presented at the 2013 New Techniques and Technologies for Statistics conference (NTTS), Brussels, Belgium, 2013. 21 December.
17 Day, H.R., Parker, J.D. Self-report of diabetes and claims-based identification of diabetes among Medicare beneficiaries // National Health Statistics Reports. 2013. 1 November. No. 69; Centers for Disease Control and Prevention. Washington, DC.
18 Dinan, K. Local Agency lessons on implementing random assignment: An example from NYC's Child Support Program / Presented at the Association for Public Policy Analysis and Management (APPAM) Annual Fall Research Conference, Washington, DC, 2013. 8 November.
19 Donohue, J.J., Wolfers, J. Uses and abuses of empirical evidence in the death penalty debate // Stanford Law Review. 2006. Vol. 58. No. 3. P. 791-846.
20 Duncan, G.T., Elliot, M., Salazar-Gonzalez, J.J. Statistical confidentiality, principles and practice. New York: Springer, 2011.
21 Duong, T., Millman, S. Behavioral data as a complement to mobile survey data in measuring effectiveness of mobile ad campaign / Presented at the CASRO Digital Research Conference 2014. [Online] <http://bit.ly/1v4fFBc>. [Retrieved January 28, 2015]
22 Dwork, C. A firm foundation for private data analysis // Communications of the ACM. 2011. Vol. 54. No. 1. P. 86-95.
23 Evans, D.S. Tests of alternative theories of firm growth // The Journal of Political Economy. 1987. Vol. 95. No. 4. P. 657-674.
24 Executive Office of the President. Big Data: seizing opportunities, preserving values. Washington DC, 2014. [Online] <http://1.usa.gov/1hqgibM>. [Retrieved January 28, 2015]
25 Fan, J., Han F., Liu, H. Challenges of Big Data analysis // National Science Review. 2014. Vol. 1. P. 293-314.
26 Fan, J., Liao, Y. Endogeneity in ultrahigh dimension // Annals of Statistics. 2014. Vol. 42. No. 3. P. 872-917.
27 Fan, J., Samworth, R., Wu, Y. Ultrahigh dimensional feature selection: beyond the linear model // The Journal of Machine Learning Research. 2009. Vol. 10. P. 2013-2038.
28 Gelman, A., Fagan, J., Kiss, A. "An analysis of the new york city police department's "stop-and-frisk" policy in the context of claims of racial bias // Journal of the American Statistical Association. 2007. Vol. 102 No. 479. P. 813-823.
29 Gerber, E.R."The Privacy context of survey response: An ethnographic account // Confidentiality, disclosure and data access: Theory and practical applications for statistical agencies / Edited by P. Doyle, J. Lane, J. Theeuwes, L. Zayatz. Amsterdam: Elsevier, 2001. P. 371-395.
30 Greenwood, D., Stopczynski, A., Sweatt, B., Hardjono, T., Pentland, P. The new deal on data: A framework for institutional controls // Privacy, Big Data, and the public good: Frameworks for engagement / Ed. by J. Lane, V.

Stodden, S. Bender, and H. Nissenbaum. Cambridge: Cambridge University Press, 2014. P. 192-200.
31 Griffin, J. The role of the Chief Data Officer // DM REVIEW. 2008. Vol. 18. No. 2. P. 28.
32 Groves, R.M. "Three eras of survey research // Public Opinion Quarterly. 2011a. Vol. 75. No. 5. P. 861-871.
33 Groves, R.M. "Designed Data" and "Organic Data." Directors Blog, 2011b. 31 May. U.S. Census Bureau. [Online] <http://1.usa.gov/15NDn8w>. [Retrieved 28 January, 2015]
34 Halevy, A., Norvig, P., Pereira, P. The unreasonable effectiveness of data // Intelligent Systems, IEEE, 2009. Vol. 24. No. 2. P. 8-12.
35 Hall, P., Miller, H. Using generalized correlation to effect variable selection in very high dimensional problems // Journal of Computational Graphical Statistics. 2009. Vol. 18. No.3. P. 533-550.
36 Hey, T., Tansley, S., Tolle, K. The fourth paradigm: Data intensive scientific discovery. New York: Microsoft Research, 2009.
37 Ibrahim, J., and Chen, M.-H. Power prior distributions for regression model // Statistical Science. 2000. Vol. 15. No. 1. P. 46-60.
38 Jovanovic, B. Selection and the evolution of Industry // Econometrica: Journal of the Econometric Society. 1982. Vol. 50. No. 3. P. 649-670.
39 Kahneman, D. Thinking fast and slow. New York: Farrar, Straus and Giroux, 2011.
40 Karr, A.F., Reiter, J.P. Using statistics to protect privacy // Privacy, Big Data, and the public good: Frameworks for engagement / Ed. by J. Lane, V. Stodden, S. Bender, and H. Nissenbaum. Cambridge: Cambridge University Press, 2014. P. 276-295.
41 Keller, S.A., Koonin, S.E., Shipp, S. Big Data and city living – What can it do for us? // Significance. 2012. Vol. 9. No. 4. P. 4-7.
42 Kinney, S.K., Karr, A.F., Gonzalez Jr, J.F. Data confidentiality: The next five years summary and guide to papers // Journal of Privacy and Confidentiality. 2009. Vol. 1. No. 2. P. 125-134.
43 Koonin, S.E., Holland, M.J. The value of big data for urban science // Privacy, Big Data, and the public good: Frameworks for engagement / Ed. by J. Lane, V. Stodden, S. Bender, and H. Nissenbaum. Cambridge: Cambridge University Press, 2014. P. 137-152.
44 Kreuter, F., Peng, R.D. Extracting Information from Big Data: Issues of measurement, inference and linkage // Privacy, Big Data, and the public good: Frameworks for engagement / Ed. by J. Lane, V. Stodden, S. Bender, and H. Nissenbaum. Cambridge: Cambridge University Press, 2014. P. 257-275.
45 Lane, J., Stodden, V. What? Me worry? What to do about privacy, big data, and statistical research // AMStat News, 2013. 1 December. [Online] <http://bit.ly/15UL9OW>. [Retrieved 28 January 2015]
46 Lane, J., Stodden, V. Bender, S., Nissenbaum, H. 2014. Editors'// Privacy, Big Data, and the public good: Frameworks for engagement / Ed. by J. Lane, V. Stodden, S. Bender, and H. Nissenbaum. Cambridge: Cambridge University Press, 2014. P. Pp. xi-xix.
47 Laney, D. "3-D data management: Controlling data volume, velocity and variety // META Group Research Note, 2001. 6 February. [Online]

<http://gtnr.it/1bKfIKH>. [Retrieved 28 January 2015]
48 Laney, D. The importance of 'Big Data': A definition. New York: Gartner Inc., 2012.
49 Lazer, D.M., Kennedy, R., King, G., Vespignani, A. The parable of google flu: Traps in big data analysis // Science. 2014. No. 343 (6176). P. 1203-1205.
50 Leek, J. Why big data is in trouble: they forgot about applied statistics // Simplystats blog. 2014a. 7 May. [Online] <http://bit.ly/1fUzZO1> [Retrieved 28 January 2015]
51 Leek, J. 10 things statistics taught us about big data analysis // Simplystats blog, 2014b. 22 May. [Online] <http://bit.ly/S1ma4Z>. [Retrieved 28 January 2015]
52 Levitt, S.D., Miles, T.J. Economic contributions to the understanding of crime // Annual Review of Law and Social Science. 2006. Vol. 2. P. 147-164.
53 Lohr, S. The Age of Big Data // New York Times. 2012. 11 February. [Online] <http://nyti.ms/1f7WKqh>. [Retrieved 28 January 2015]
54 Lohr, S. For Big-Data scientists, 'janitor work' is key hurdle to insights // New York Times. 2014. 17 August. [Online] <http://nyti.ms/1Aqif2X>. [Retrieved 28 January 2015]
55 Manzi, J. Uncontrolled: The surprising payoff of trial-and-error for business, politics, and society. New York: Basic Books, 2012.
56 McAfee, A., Brynjolfsson, E. Big Data: The management revolution // Harvard Business Review. 2012. Vol. 90. No. 10. P. 61-67.
57 Murphy, J., Link, M.W., Childs, J.H., Tesfaye, C.L., Dean, E., Stern, M., Pasek, J., Cohen, J., Callegaro, M., Harwood, P. Social media in public opinion research: Report of the AAPOR Task Force on Emerging Technologies in Public Opinion Research. AAPOR Task Force Report. 2014. [Online] <http://bit.ly/15V7coJ>. [Retrieved 28 January 2015]
58 New York Times Editorial Board. Better governing through data // New York Times. 2014. 19 August. [Online] <http://nyti.ms/1qehhWr>. [Retrieved 28 January 2015]
59 Nielsen, Michael. 2012. Reinventing Discovery: The New Era of Networked Science. Princeton, NJ: Princeton University Press.
60 Nissenbaum, H. A contextual approach to privacy online // Daedalus. 2011. Vol. 140. No. 4. P. 32-48.
61 Norberg, A., Sammar, M., Tongur, C. A study on scanner data in the Swedish Consumer Price Index / Presentation to the Statistics Sweden Consumer Price Index Board. Stockholm, Sweden, 2011. 10-12 May.
62 Ohm, P. Broken promises of privacy: Responding to the surprising failure of anonymization // UCLA Law Review. 2010. Vol. 57. No. 6. P. 1701-1818.
63 Pardo, T.A. Making data more available and usable: A getting started Guide for Public Officials / Presentation at the Privacy, Big Data and the Public Good Book Launch. 2014. 16 June. [Online] <http://bit.ly/1Czw7u4>. [Retrieved 28 January 2015]
64 Porter, S., Lazaro, C.G. Adding Big Data booster packs to survey data / Presented at the CASRO Digital Research Conference. San Antonio, TX, 2014. 12 March.
65 Prada, S.I., González-Martínez, C., Borton, J., Fernandes-Huessy, J., Holden,

C., Hair, E., Mulcahy, T. Avoiding disclosure of individually identifiable health information. A Literature Review. London: SAGE Open, 2011. [doi:10.1177/2158244011431279]
66 Schenker, N., Davidian, M., Rodriguez, R. The ASA and Big Data // AMStat News. 2013. 1 June. [Online] <http://bit.ly/15XAzX8>. [Retrieved 28 January 2015]
67 Scott, S.L., Blocker, A.W., Bonassi, F.V., Chipman, H.A., George, E.I., McCulloch, R.E.. Bayes and Big Data: The consensus Monte Carlo algorithm. 2013. [Online] <http://bit.ly/1wBqh4w>. [Retrieved 28 January 2015]
68 Shelton, T., Poorthuis, A., Graham, M., Zook, M. Mapping the data shadows of hurricane sandy: uncovering the sociospatial dimensions of 'Big Data' // Geoforum. 2014. Vol. 52. P. 167-179.
69 Squire, P. Why the 1936 Literary Digest Poll Failed // Public Opinion Quarterly. 1988. Vol. 52. No. 1. P. 125-133.
70 Stanton, M.W. The high concentration of U.S. Health Care Expenditures // Research in Action. 2006. Vol. 19. P. 1-10.
71 Stock, J.H., Watson, M.W. Forecasting using principal components from a large number of predictors // Journal of the American Statistical Association. 2002. Vol. 97. No. 460. P. 1167-1179.
72 Strandburg, K.J. 2014. Monitoring, datafication, and consent: Legal approaches to privacy in the Big Data Context." // Privacy, Big Data, and the public good: Frameworks for engagement / Ed. by J. Lane, V. Stodden, S. Bender, and H. Nissenbaum. Cambridge: Cambridge University Press, 2014. P. 5-43.
73 Tambe, P., Hitt, L.M. The productivity of information technology investments: New evidence from IT labor data // Information Systems Research. 2012. Vol. 23. P. 599-617.
74 Tapia, A.H., LaLone, N., Kim, H.-W. Run Amok: Group crowd participation in identifying the Bomb and Bomber from the Boston Marathon Bombing / Proceedings of the 11th Int. ISCRAM Conference. Pennsylvania, PA. 2014.
75 Taylor, S.J. Real scientists make their own data // Sean J Taylor Blog, 2013. 25 January. [Online] <http://bit.ly/15XAq5X>. [Retrieved 28 January 2015]
76 Thompson, W.W., Comanor, L., Shay, D.K. Epidemiology of seasonal influenza: use of surveillance data and statistical models to estimate the burden of disease // Journal of Infectious Diseases. 2006. Vol. 194. (Supplement 2). P. S82-S91.
77 Tourangeau, R. Rips, L.J., Rasinski, K. The psychology of survey response. Cambridge: Cambridge University Press, 2000.
78 Varian, H.R. Big Data: New tricks for econometrics // The Journal of Economic Perspectives. 2014. Vol. 28. No. 2. P. 3-27.
79 Wallgren, A., Wallgren, B. Register-based statistics: Administrative data for statistical purposes. New York: Wiley & Sons, 2007.
80 Wallgren, A., Wallgren, B. Register-based statistics: Statistical methods for administrative data. New York: Wiley & Sons, 2014.
81 Winkler, W.E. "Re-Identification methods for evaluating the confidentiality of analytically valid microdata / Research Report Series, Statistics #2005-09. U.S. Census Bureau. 2005.

10. Терминологический словарь

Большие данные (Big Data): Данные, настолько большие по объему, что обработка этих данных затруднительна. С данными может быть трудно работать ввиду их размера (объем) и / или скорости, с которой они генерируются (скорость), и / или формата, в котором они генерируются, как, например, текстовые документы или изображения (многообразие).

Естественные (или органические) данные (Found data): Данные, созданные как побочный результат другого процесса или деятельности (например, данные, полученные с сенсорных устройств на конвейере, или временные метки и геоданные, созданные в твиттере).

Сконструированные данные (Made data): Или созданные данные (designed data). Данные, которые создаются с четко определенной целью (например, данные опросов или данные эксперимента).

Структурированные данные (Structured data): Числовые и категориальные данные, которые подходят для традиционных реляционных баз данных. Большая часть данных, с которыми «комфортно» работать, могут рассматриваться как структурированные данные.

Неструктурированные данные (Unstructured data): Данные, в которых не прослеживается четкая структура (например, текст PDF файлов, видеозаписи с камер наблюдения), и потому для работы с ними требуют дополнительной обработки и структурирования.

Процесс генерации данных (Data-generating process): Или функция вероятностей. Процесс, во время которого генерируются данные (то есть откуда берутся данные).

Объектно-ориентированная платформа расширенного доступа к данным (Hadoop): открытая распределенная файловая система, которая может хранить как структурированные, так и не структурированные данные. В дополнение к этому все данные дублируются, так что даже при выходе из строя аппаратного оборудования никакие данные не утрачиваются.

Модель распределённых вычислений (Map-Reduce): Парадигма обработки данных по принципу «разделяй и властвуй», когда сложные вычислительные операции распределяются между несколькими компьютерами, что сокращает общее время вычислительного процесса (например, если 10 компьютеров обрабатывают 1 миллиард запросов, то это занимает меньше времени, чем если бы один компьютер обрабатывал 10 миллиардов).

МАТЕРИАЛЫ V СОЦИОЛОГИЧЕСКОЙ ГРУШИНСКОЙ КОНФЕРЕНЦИИ

Научное издание

ОТЧЁТ AAPOR О БОЛЬШИХ ДАННЫХ 12 ФЕВРАЛЯ 2015

Подписано в печать .2015 г. Формат70х100/16.
Печать офсетная. Усл. печ. л. 7,75
Тираж 500 экз. Заказ № 40-02/15

Отпечатано в ООО «Центр полиграфических услуг «Радуга»»
Тел.: (495) 739-56-80
http://www.raduga-print.ru/
http://www.radugaprint.ru/