# ПЕРВАЯ СТЕПЕНЬ

**И.Е. КОЧЕДЫКОВ***

## ОБ ОПЫТЕ ПРИМЕНЕНИЯ БОЛЬШИХ ДАННЫХ В ПОЛИТИЧЕСКОЙ НАУКЕ

*Аннотация*. Данная обзорная статья посвящена анализу успехов и проблем применения больших данных в политической науке. В первой части обсуждаются онтологические эпистемологические предпосылки использования Big data и машинного обучения в политических исследованиях. Во второй части автор проводит обзор репрезентативных результатов политологических исследований, произведенных с использованием больших данных. Третья часть статьи посвящена критике и определению пределов использования Big data в политических исследованиях. Автор показывает, что кроме чисто технических проблем, связанных, например, с неполнотой имеющихся данных, искажениями из-за присутствия ботов, существует серьезные ограничения возможностей применения больших данных для анализа политических действий, которые имеют диспозициональный характер.

*Ключевые слова:* Big data; большие данные; алгоритмы; машинное обучение; онтология и эпистемология Больших данных; теория действия.

*Для цитирования:* Кочедыков И.Е. Об опыте применения больших данных в политической науке // Политическая наука. – 2023. – № 4. – С. 226–251. – DOI: http://www.doi.org/10.31249/poln/2023.04.09

Большие данные и машинное обучения за последние десять лет получили некоторое распространение в политической науке.

* **Кочедыков Иван Евгеньевич**, магистр, аспирант, Национальный исследовательский университет «Высшая школа экономики» (Москва, Россия), e-mail: Kochedykov.ivan@gmail.com

 DOI: 10.31249/poln/2023.04.09

Этому способствовало увеличение производительной мощности компьютеров, появление множества новых источников данных в коммерческом секторе (речь идет, прежде всего, о социальных сетях), рост доступности собираемых государством данных в некоторых странах, а также успехи в создании новых алгоритмов анализа данных [Grossman, 2020, p. 227]. Некоторые исследователи стали видеть в анализе таких данных решение почти всех исследовательских проблем, возможность точного предсказания трендов и выработки политических рекомендаций. И хотя политологи и социологи изначально довольно скептично, по сравнению со специалистами из компьютерных наук, отнеслись к появлению нового типа данных, они постепенно нашли им применение в целом ряде тематических исследований, как то: предсказания результатов выборов, анализ общественных предпочтений, изучение эффектов определенных политических курсов, автоматизация рутинных операций и т.д. [Grossman, 2020].

В данной статье не будут затрагиваться вопросы политических последствий применения больших данных, будь то влияние на выборы, поляризация общества и т.д. Внимание автора сосредоточено исключительно на вкладе Big data в научные исследования. В первой части статьи речь пойдет об онтологии и эпистемологии Big data. Во второй части будут кратко рассмотрены основные примеры применения больших данных для анализа политических процессов, событий, результатов. Исследования для иллюстрации отбирались по критериям цитируемости и вхождения в различные учебные программы. Естественно, что представленные примеры не являются исчерпывающими. В третьей части статьи мы рассмотрим основные критические возражения, касающиеся данной группы методов. В завершении обсуждаются перспективы использования больших данных в политической науке.

## Онтология больших данных

Несмотря на то, что термин Big data уже довольно давно существует в исследованиях и медиасреде, он так и не получил однозначного определения [Salganik, 2019, p. 14]. Предполагается, что в

Big data входят данные онлайн-платформ (логи[1], посты в социальных сетях), коммерческие данные компаний (например, логи производства, покупок), собираемые государством данные (статистические записи, платежи, налоги, штрафы и т.д.)

Трактовка больших данных французским философом и социологом Д. Булье хотя и не получила пока широкого распространения, вызывает большой интерес. Автор полагает, что большие данные как новый технологический феномен переопределяют онтологию социального, и, соответственно, методологию его исследования [Boullier, 2015]. Базовой онтологической категорией становится категория «след» (trace), а не личности (person), идентичности или сообщества. Цифровые следы – отпечатки активности (не обязательно только человека) в цифровом пространстве. Следы, по мнению французского ученого, могут варьироваться от сигналов («сырых», генерируемых объектами) до неструктурированных стенограмм, распространяющихся в виде мемов (или цитат); это могут быть метаданные, ссылки, клики, лайки, куки [Boullier, 2015, p. 10]. Булье утверждает, что, анализируя новую реальность, необходимо отслеживать алгоритмы, которые ориентируются только на те или иные свойства в зависимости от предзаданных целей. При этом данные свойства не являются структурными (что характерно было для «старой» социальной науки), Булье их называет «вторичными» и полагает, что они основываются на аппроксимациях, скорректированных в результате обучения на больших объемах данных (например, предыдущая покупка или лайк на сообщение). Следы, понимаемые в этом строгом смысле, порождаются цифровыми платформами и технологическими системами, но не являются «знаками» или индикаторами чего-либо, кроме самих себя и (возможных) отношений с другими атрибутами.

Следы, по мнению французского ученого, в принципе независимы от других атрибутов, особенно социально-демографических, которые редко мобилизуются в искомых корреляциях между следами (что, опять же, отличает новую цифровую реальность от свойственных классической социальной науке). Связи с наиболее распространенными в науке о данных параметрами ограничиваются

[1] Файл с записями о событиях в хронологическом порядке, простейшее средство обеспечения журналирования.

временем (временными отметками о появлении следа) и местом (геолокационные метки), что позволяет создавать временные шкалы и карты, которые становятся способами упрощенного представления следов. На основе этих общих ссылок можно проводить корреляции между всеми типами данных. Однако благодаря такому радикальному упрощению, когда в качестве референтных сущностей уже не выступают индивиды со всеми их социально-демографическими свойствами, следы быстро циркулируют и изменяют состояние самих баз данных, которые становятся динамичными.

Более приземленный подход дает известный специалист и пионер Big data Р. Китчин. Он обобщил существующие определения и выделил семь признаков больших данных [Kitchin, 2016]. В частности, большие данные характеризуются такими аспектами, как объем (охват огромного количества данных за счет технических возможностей хранения), скорость (создание фактических данных в режиме реального времени), разнообразие (интеграция различных подходов к структурированию), исчерпываемость (захват полной выборки), разрешение и идентификация (разная степень детализации и обращение к множеству сущностей и процессов), реляционность (связь между несколькими наборами данных), расширяемость и масштабируемость (добавление новых данных и увеличение объема), достоверность (сохранение неопределенности и ошибок). По получившему широкое распространение мнению канадских исследователей, к ним необходимо добавить ценность (извлечение множества различных выводов) и изменчивость (изменение смысла в зависимости от контекста) [Gandomi, 2015].

По мнению израильского политолога Дж. Гроссмана, из всех этих признаков для политологии наиболее актуальны два – разнообразие данных и технологические средства, необходимые для их извлечения, организации и анализа [Grossman, 2020, p. 232]. Таким образом, в политической науке наиболее простым и релевантным определением будет следующее: большие данные – это как структурированные, так и неструктурированные данные различного происхождения, к которым можно получить доступ, проанализировать и обработать с помощью цифровых технологий [Grossman, 2020, p. 232–233].

Здесь также необходимо остановиться и на машинном обучении как составной части Big data. Машинное обучение (и глубокое обучение как его подобласть) – это область знаний-навыков,

которая находится на пересечении статистики и компьютерной науки. Она определяется как использование статистических методов, позволяющих компьютерным системам постепенно улучшать свою работу над конкретной задачей, используя данные без явного программирования [Chatsiou, 2020, p. 4]. Подразумевается, что компьютерная программа может начинать с исходной модели данных, анализировать фактические данные, учиться на основе этого анализа и автоматически обновлять исходную модель, чтобы учесть результаты анализа.

## Эпистемология больших данных

Несмотря на то что некоторые эпистемологические подходы к работе с большими данными уже частично проговаривались выше в описании их онтологических характеристик, есть необходимость рассмотреть их подробнее. Один из ведущих специалистов в этой области, сотрудник Мюнхенского центра технологий в обществе (Munich Center for Technology in Society) при Техническом университете Мюнхена Вольфганг Питш (Wolfgang Pietsch), полагает, что наука о данных и ее применение в разных предметных областях представляет собой индуктивистский подход к получению знаний [Pietsch, 2021, pp. 38–39]. То есть путь к знанию идет от частного к общему, что отличается от более традиционных подходов количественных политических исследований, где разработка наборов теоретических моделей предваряет изучение эмпирических объектов.

Наука о данных начинается с огромных объемов информации. Под данными понимаются записи эмпирических свидетельств того или иного явления, т.е. зафиксированные наблюдения или результаты экспериментов. Например, данные, используемые поисковой системой, могут включать предыдущие поисковые запросы или различные характеристики пользователя, такие как возраст, пол, национальность и т.д. Все эти данные представляют собой конкретные факты, а не общие законы или гипотезы [Pietsch, 2021, p. 44–45].

Помимо данных, вторым важнейшим компонентом науки о данных являются различные алгоритмы машинного обучения, способные делать предсказания на основе данных, т.е. индуктивно

выносить умозаключения о пока неизвестных событиях или лицах. Так, поисковая система обычно делает индуктивный вывод на основе всех релевантных данных, какой ответ лучше всего подходит для конкретного запроса пользователя. Хотя работа этих алгоритмов иногда может быть непрозрачной из-за огромного количества выполняемых шагов, в том, как они достигают своих результатов, нет ничего мистического. В частности, здесь мало места для интуиции или творчества, по крайней мере после того, как получены данные и выбраны алгоритмы. Принципы работы большинства алгоритмов машинного обучения хорошо понятны, и после настройки эти алгоритмы редко нуждаются в дальнейшем вмешательстве человека. Примером этого, по мнению Питша, является тот факт, что ответы на поисковые запросы обычно генерируются без участия человека.

Концептуальное ядро каузальности в науке о больших данных заключается в том, что изменение обстоятельств приводит к изменению исследуемых явлений [Pietsch, 2021, p. 73–80]. Питш предлагает понимать причинность Big data в терминах различий (как у Дж.С. Милля). Причинно-следственные связи определяют факторы различия в том смысле, что если изменить определенную переменную, то это окажет влияние на другую переменную. То есть причинно-следственные связи, выявленные в ходе исследования, не являются объяснительными в сильном смысле этого слова. Наука о данных дает объяснения, указывая причины или, по крайней мере, их приближенные варианты, но при этом не ссылается на некие объединяющие принципы, которые характерны для гипотетико-дедуктивных подходов. При этом необходимо иметь в виду, что индуктивистские концепции часто нацелены на предсказание явлений, а не на их фундаментальное понимание.

Что касается проблемы истины, то в феноменологической науке (какой является, по мнению Питша, наука о больших данных) понятие истины доступно в терминах соответствия причинных законов представленному через данные миру в ощущениях. Обратной стороной данного подхода является контекстуальный и приблизительный характер феноменологических законов по сравнению с амбициями универсальности и точности теоретической науки [Pietsch, 2021, p. 97–98].

Для науки о данных в плане получения знаний характерно непараметрическое моделирование [Pietsch, 2015]. Оно предпола-

гает малое количество модельных допущений, например таких, как широкий спектр функциональных зависимостей или функций распределения. При непараметрическом моделировании прогнозы рассчитываются на основе «всех» данных. Хотя это делает непараметрическое моделирование достаточно гибким, позволяя быстро реагировать на неожиданные данные, оно также становится чрезвычайно требовательным к объему данных и вычислений.

В исследованиях с большими данными применяют три основных класса алгоритмов [Ын, Су, 2022]. Первый класс – «Обучение без учителя» (unsupervised learning). Его применяют для нахождения скрытых закономерностей в наборе, когда мы не знаем, какие закономерности искать. В этот класс входят такие алгоритмы, как a) метод k-средних, b) метод главных компонент, c) ассоциативные правила и d) анализ социальных сетей. Второй класс – обучение с учителем (supervised learning). Он используется для прогнозирования заданных шаблонов, например проверки точности модели. Здесь используются такие методы, как a) регрессионный анализ, b) метод k-ближайших соседей, c) метод опорных векторов, d) древо решений, e) случайные леса, f) создание и тренировка нейросетей. Третий класс – обучение с подкреплением, которое использует закономерности в данных для улучшения прогнозирования по мере появления новых результатов. Главным представителем является алгоритм «Многорукие бандиты»[1].

## Применение Big data в политологических исследованиях

### *а) Прогнозирование электоральных результатов*

Распространение социальных сетей и возросшая активность пользователей сделала их привлекательным источником данных для политических исследований и прогнозирования. Исследователи исходят из того, что анализ динамики избирательной кампании с помощью данных социальных сетей и их связи с результатами голосования имеет ряд очевидных преимуществ. Будучи более дешевым и быстрым по сравнению с традиционными опросами

[1] См. также: Кто же такой этот многорукий бандит? – Хабр. – 21 сентября 2022. – Режим доступа: https://habr.com/ru/articles/689364/ (дата посещения: 01.07.2023).

общественного мнения, он позволяет прогнозировать ход кампании, т.е. отслеживать в реальном времени (день за днем или даже час за часом) эволюцию предпочтений, чтобы уловить тенденции и любые (возможные) резкие изменения в общественном мнении быстрее, чем традиционные опросы [Ceron, 2017].

Для прогнозирования электоральных результатов исследователи анализируют огромное количество интернет-источников, начиная от блогов и социальных сетей, таких, как запрещенные в России Facebook и Twitter, онлайн-новостей, поиска Google и данных о просмотрах страниц Википедии. В зависимости от метода, используемого для прогнозирования, существующие исследования можно разделить на три основных подхода. Первый подход является чисто количественным и опирается на автоматизированные вычислительные методы подсчета данных. Примером может служить исследование индийских социальных сетей в 2014 г. [Barclay, 2015]. Ученые использовали количество «лайков» (отметок «нравится») на официальных страницах кандидатов для оценки их популярности в контексте выборов в Лок Сабха. Они обнаружили сильную положительную корреляция между количеством «лайков» и долей голосов избирателей. Исследователи также установили, что месяц, предшествующий периоду голосования, является наилучшим для прогнозирования доли голосов с помощью отметок «нравится» – с точностью 86,6%.

Второй подход обращает внимание на язык и пытается придать качественное значение комментариям (постам, твитам), публикуемым пользователями социальных сетей с помощью использования автоматизированных инструментов для анализа тональности (sentiment analysis). Нидерландские исследователи Эрик Тьонг Ким Санг и Йохан Бос использовали данные Twitter для предсказания результатов выборов в сенат страны в 2021 г. [Kim, Bos, 2012]. Авторы создали корпус политических твитов с помощью ручной аннотации тональностей. Они вручную аннотировали 1678 политических твитов, присваивая каждому твиту один из двух классов: отрицательный по отношению к партии, упомянутой в твите, или неотрицательный. Распределение мест между партиями, предсказанное твитами, оказалось близким к результатам выборов.

Третий метод строится во многом на тех же основаниях, но использует полуавтоматический анализ тональности с учителем

для выявления (агрегированного) мнения, выраженного в интернете [Grimmer, 2013]. Данный метод основан на двухэтапном процессе, где на первом этапе исследователи читают и кодируют подвыборку документов. Она представляет собой обучающее множество, которое на втором этапе алгоритма будет использоваться для классификации всех непрочитанных документов (тестовое множество). На втором этапе агрегированная статистическая оценка алгоритмов анализа тональности с учителем распространяет выявленную точность на всю совокупность постов, позволяя корректно получать мнения, высказываемые в социальных сетях.

В качестве иллюстрации данного подхода обратимся к выполненному итальянскими политологами анализу президентской кампании в США 2012 г. и двум турам первичных выборов, проведенных итальянской левоцентристской коалицией в ноябре 2012 г. [Ceron, 2017]. В целом произведенный анализ социальных сетей позволил правильно предсказать победителя в девяти из 11 колеблющихся штатов, исключение составляют Колорадо и Пенсильвания. Более того, в большинстве штатов (семь против двух) данные исследователей оказались точнее, чем среднее значение опросов (это Флорида, Айова, Вирджиния, Невада, Нью-Гэмпшир, Мичиган и Висконсин), а в двух оставшихся штатах (Огайо и Северная Каролина) разные прогнозы (социальные медиа против опросов) оказались одинаковыми. В случае итальянских выборов расхождения между оценками политологического коллектива и результатами волеизъявления оказались незначительны и в среднем не превышали 2%, что соответствует средней погрешности опросов. Кроме того, авторы утверждают, что их методика позволила предсказать разрыв между двумя ведущими кандидатами (Берсани и Ренци) лучше, чем традиционные опросы. Согласно их оценкам, разрыв между двумя кандидатами составлял 10,5%, в то время как Берсани после подсчета голосов лидировал на 9,4 пункта (т.е. разница в 1,1%, тогда как в среднем опросы ошибались в величине разрыва на 3%).

### *b) Предсказание конфликтов*

Сотрудники Мичиганского университета использовали машинное обучение для анализа факторов, способствующих улуч-

шению предсказания гражданских войн [Colaresi, 2017]. Авторы использовали исторические, социальные, экономические, политические и демографические данные для научения машинного интеллекта (Box's loop, случайные леса) предсказывать возможности появления гражданских конфликтов в разных регионах мира. Они пришли к выводу, что экспорт сырьевых товаров служит важной движущей силой гражданского конфликта и является полезным опережающим индикатором.

В другом исследовании для предсказания случаев политического насилия политологи провели анализ газет [Mueller, Rauh, 2018]. Они применяют тематические модели (topic model), которые позволяют уменьшить размерность текста с количества, близкого к миллиону выражений, до, например, 15 тем. Эти темы затем были использованы в простых линейных регрессионных моделях для предсказания начала конфликта.

Метод, полностью основанный на темах, способен довольно хорошо предсказать время возникновения конфликта. Во-первых, результаты можно легко интерпретировать, поскольку темы представляют собой содержательные резюме текста. Во-вторых, алгоритм, генерирующий темы, способен учиться на изменяющихся ассоциациях терминов. Например, в статье авторы показали, что новые термины, такие как «террорист» или «повстанец», служат ключевыми индикаторами риска конфликта в последние годы, в то время как в 1995 г. они таковыми не являлись. В-третьих, тематическая модель использует в прогнозе отрицательные ассоциации между темами и конфликтным риском. На самом деле, значительная часть прогноза, по-видимому, исходит от тем, не имеющих прямого отношения к конфликту. Особенно сильна связь между меньшим количеством сообщений о судебных процедурах и правоохранительной деятельности и более высоким риском конфликтов.

### *c) Анализ политических предпочтений, взглядов и мотиваций*

Тематическое моделирование также используется для анализа политических предпочтений, позиций элит. В частности, в статье [Bonica, 2016] описывается трехэтапная стратегия моделирования для измерения предпочтений и выраженных приоритетов по различным политическим аспектам. Она сочетает в себе тематическое

моделирование, оценку идеальной точки и методы машинного обучения. На первом этапе к базе данных политических текстов автор применил тематическое моделирование. На втором этапе были расставлены идеальные баллы законодателей по конкретным вопросам на основе данных о голосовании в прошлом с использованием оцененных весов тем для определения размерности списков. На третьем этапе было проведено обучение с помощью метода опорных векторов[1] для прогнозирования баллов по вопросам для более широкого круга кандидатов на основе общих источников данных (например, данных о пожертвованиях). Результатом стало создание единого ресурса данных об американской политической элите.

В статье [Bond, Messing, 2015] представлен матричный метод измерения идеологии с использованием масштабных данных из социальных сетей. В частности, в исследовании рассматривается политическая поляризация путем наложения идеологии на структуру социальных отношений. Авторы выяснили, что дружеские связи имеют тенденцию к кластеризации на основе идеологических предпочтений. Кроме того, в статье показано, что идеологическая корреляция сильнее проявляется среди близких друзей, членов семьи и особенно романтических партнеров. Индивиды, находящиеся в сетях с сильными идеологическими разногласиями, менее склонны к участию в политике.

В работе отечественных исследований [Ахременко, Петров, 2023] была проанализирована мотивация участников белорусских протестов в 2020 г. по данным социальной сети VK. Авторы с помощью 12 кодировщиков соотнесли публикации, получившие наибольшее количество репостов, с тремя ключевыми мотивациями, стоящими за участием в протестах (гнев, идентификация с протестным движением, вера в успех коллективного действия). Полученные динамические ряды политологи сравнили с развитием уличной протестной активности и пришли к выводу, что она спала, когда протестующие потеряли веру в успех своих выступлений.

Развитие методов работы с большими данными недавно привело к возможности перейти от анализа отдельных слов (что

[1] См. также: Краткий обзор алгоритма машинного обучения Метод Опорных Векторов (SVM). – Режим доступа: https://habr.com/ru/articles/428503/ (дата посещения: 03.07.2023).

характерно для тематического моделирования и анализа тональности) к анализу нарративов [Ash, 2023]. Отправной точкой метода поиска нарративов является семантическая маркировка ролей (semantic role labeling, SRL) – лингвистический алгоритм, который, принимая предложение в виде обычного текста, определяет действие, агента, выполняющего это действие, и пациента, на которого оно направлено. Полученное пространство признаков агентов, действий и претерпевающих действие гораздо более информативно для нарративов, чем пространство признаков, создаваемое стандартными методами «текст как данные».

Следующей частью описываемого метода поиска нарративов является набор процедур снижения размерности. Данный подход к кластеризации сущностей использует несколько вариантов фраз, относящихся к одному и тому же объекту (например, «налоги на доходы» и «налогообложение доходов» или «бывший президент Рейган» и «Рональд Рейган»), и сводит их к одной маркировке (labeling) объекта. Авторы используют методы обучения без учителя, такие, как тематические модели и вложение документов (document embeddings)[1].

Семантическая маркировка ролей – это алгоритм вычислительной лингвистики, который отвечает на основные вопросы – кто действующее лицо события, например грамматический субъект глагола в активном залоге письменных предложений, в частности, кто и что делает с кем. Глагол («что делает») фиксирует действие в залоге. Претерпевающий действие («кого, кому, кем») – это субъект, на которого распространяется действие, т.е. объект или цель. Семантические роли не только различают действия и сущности в предложении, но и отражают отношения между ними. Например, SRL извлекает ту же направленную связь для предложения «Миллионы американцев потеряли свои пособия по безработице», что и для инвертированного предложения «Пособия по безработице потеряли миллионы американцев».

Применение данного метода для анализа речей американских конгрессменов показало, что к наиболее позитивным нарративам относятся нарративы о Конституции и отцах-основателях, о

[1] См. также: Чудесный мир Word Embeddings: какие они бывают и зачем нужны? – Режим доступа: https://habr.com/ru/companies/ods/articles/329410/ (дата посещения: 03.07.2023).

преимуществах здравоохранения, о малом бизнесе, обеспечивающем рабочие места. Негативный набор включает высказывания о предоставлении помощи в трудные времена и об угрозах, исходящих от террористов. В итоге результаты позволили получить качественное представление о приоритетах и ценностях, которыми руководствуются конгрессмены США, а также об их идеологических разногласиях.

*d) Анализ политических коммуникаций*

Большое внимание в новых исследованиях уделяется политической коммуникации. Так, в работе М. Коновера описывается структура и динамика массовой политической мобилизации и коммуникации [Conover, 2013]. С этой целью была проанализирована сеть онлайн-коммуникаций, реконструированная на основе более 600 тыс. твитов за 36 недель, охватывающих период зарождения и становления американского антикапиталистического движения «Оккупируй Уолл-стрит». Было обнаружено, что по сравнению с сетью стабильной внутриполитической коммуникации, сеть «Оккупируй Уолл-стрит» демонстрирует более высокий уровень локальности и структуру «ступицы и спицы», в которой большая часть нелокального внимания направлена на такие резонансные места, как Нью-Йорк, Калифорния и Вашингтон. Более того, информационные потоки через границы штатов чаще содержат формулировки фреймов и ссылки на СМИ, в то время как коммуникация между людьми в одном штате чаще связана с акциями протеста и конкретными местами и временем. Автор полагает, что эти особенности отражают усилия движения по мобилизации ресурсов на местном уровне и разработке нарративных фреймов, укрепляющих коллективную цель на национальном уровне.

Продолжая американскую тему, местные политологи использовали алгоритм случайных лесов для анализа динамики «оглупления» дискурса [Benoit, Munger, Spirling, 2019]. Для этого они собрали базу текстов «О положении дел в стране» (State of the union address). Затем, с помощью краудсорсинга, авторы провели тысячи парных сравнений фрагментов текста и включили полученные результаты в статистическую модель сложности, в которую также вошли такие признаки, как части речи и мера редкости

слов, полученные на основе динамических частот терминов в наборе данных Google Books. В итоге они пришли к выводу, что в последние годы обращения стали более простыми для восприятия.

Другой популярной областью исследований, к которой применяют методы науки о данных для изучения текстовых следов политики, стал поиск и обнаружение «фейковых новостей». Например, испанские специалисты в исследовании с помощью алгоритмов k-ближайших соседей, случайных лесов, наивного байесовского анализа и метода опорных векторов смогли (по их словам) правильно обнаружить почти все фейки в своей выборке данных [Reis, 2019]. Однако при этом 40% правдивых новостей были неправильно классифицированы. Так, определенная история считалась ложной, поскольку была опубликована новой газетой, размещенной на том же IP-адресе, что и известный источник фальшивых новостей, внесенный в черный список.

В статье профессора Наньянского технологического университета в Сингапуре А. Катагири и доцента Калифорнийского университета в Лос-Анджелесе Э. Мин обсуждается проблема эффективности разных типов сигналов в международных отношениях [Katagiri, Min, 2019]. Под сигналами понимаются заявления или действия, передающие информацию с намерением повлиять на представление получателя об отправителе. Авторы анализировали материальные действия, связанные с Берлинским кризисом[1], используя заголовки и выдержки из «Нью-Йорк таймс». Авторы должны были закодировать, сообщалось ли в каждой из 1601 отобранной статьи про кризис о потенциально затратных военных действиях, отражающих враждебность. Пять типов событий квалифицировались следующим образом: возведение стены (1); ядерные или ракетные испытания (7); сбитый самолет (1); блокада (1); задержание или остановка военных колонн и транспортов (81). Всего в этих статьях фигурирует 91 событие. В данном исследо-

[1] «Берлинский кризис» 1958–1961 гг. – международный политический кризис, связанный с решением проблемы управления столицей Германии, разделенной после окончания Второй мировой войны 1939–1945 гг. на оккупационные зоны. Началом его считается ультиматум Н.С. Хрущёва от 27 ноября 1958 г., конец – 1962–1963 гг., а пик кризиса пришелся на июнь – ноябрь 1961 г. Кризис разрешился после переговоров, избежав большой войны между двумя блоками. Берлин остался разделенным на две части, между которыми была возведена стена.

вании использовались модели статистического обучения с учителем для автоматизации большей части процесса кодирования.

Три пула необработанных текстов были преобразованы в количественные данные. Для этого сообщения / релизы сначала были разбиты на сегменты по 300 слов для более корректного анализа. Текст в каждом сегменте подвергается стандартной предварительной обработке. Она включает удаление стоп-слов, таких, как артикли и преобразование слов в лексемы (например, преобразование «говорит» и «говорил» в «говорить»). Количество оставшихся лексем в каждом сегменте подсчитывается и записывается. В итоге получается матрица «документ-термин» для каждой коллекции сообщений, где каждая строка представляет собой сегмент из 300 слов, а каждый столбец содержит информацию о том, сколько раз используется лексема. Эти количества лексем являются основными переменными, используемыми для обучения моделей и генерирования прогнозируемых значений советской решимости для каждого сегмента.

Проанализировав все имеющиеся данные, авторы пришли к выводу, что советские угрозы в приватных коммуникациях во время Берлинского кризиса острее воспринимались в Белом доме, тогда как публичные рассматривались как попутный шум.

### *е) Анализ хода электорального процесса*

Другой областью применения Больших данных и машинного обучения стал анализ электоральных махинаций [Montgomery, 2015]. Монтгомери и др. использовали байесовскую модель аддитивных деревьев регрессии (BART) (метод машинного обучения) на большом межнациональном массиве данных. Модель BART использует результаты выборов и контекстуальные факторы для выявления мошенничества только в той мере, в какой конкретные модели (например, распределение цифр или распределение цифр в сочетании с контекстуальными характеристиками) являются достоверными эмпирическими индикаторами мошенничества в международном масштабе.

В итоге авторы пришли к выводам, что политическая нестабильность оказывает существенное влияние на уровень мошенничества. Аналогичным образом, мошенничество на выборах более

распространено в странах с чрезвычайно высоким уровнем этнолингвистического разнообразия. Урбанизация, как оказалось, имеет в значительной степени отрицательную связь с авторским индикатором мошенничества. Низкий уровень плотности городского населения в наибольшей степени ассоциируется с высокими показателями по данному индикатору мошенничества, в то время как в странах с более высокой плотностью городского населения прогнозируются более низкие показатели. Слишком низкая и слишком высокая явка также служат хорошими признаками электоральных махинаций.

### *f) Анализ реализации политического курса (policy)*

Заслуженно большое внимание и успех получила статья о работе китайской интернет-цензуры, написанная под руководством одного из ведущих методологов в политической науке Г. Кинга [King, Pan, Roberts, 2013]. Авторы использовали несколько методов для анализа и понимания феномена интернет-цензуры в Китае. Во-первых, они собрали собственную массивную базу данных, включающую миллионы тематических сообщений в социальных сетях с китайской платформы микроблогов Sina Weibo[1]. Они проанализировали содержание этих сообщений, чтобы выявить в них критику правительства или упоминания о коллективных действиях. На втором этапе исследователи разработали алгоритм машинного обучения для выявления цензуры в собранном наборе данных. Алгоритм был обучен на закодированных вручную образцах цензурированных и нецензурированных сообщений с целью выявления закономерностей и индикаторов цензуры. Далее авторы провели серию рандомизированных контролируемых испытаний, чтобы понять специфику механизмов цензуры, применяемых китайским правительством. Они систематически размещали сообщения с различными типами чувствительного контента, чтобы проследить вариации цензурных практик. В итоге они пришли к выводу, что китайская политика в области регулирования интернета допускает политическую критику, но при этом негативно настроена к организации коллективных действий.

[1] Было выделено 85 тем.

*g) Исследования политического поведения*

Большие данные используются и для анализа непосредственного политического поведения. Израильские политологи А. Ротман и М. Шалев обратились к данным о местоположении пользователей мобильных телефонов для измерения мобилизации в массовых акциях протеста [Rotman, 2022]. Подобные данные не только позволили оценить количество и состав участников крупных демонстраций, но и определить, когда, где и с кем различные социально-политические слои объединяются в протестной кампании. Авторы выделили три типа сообществ, которые приняли участие в уличных протестах, а также связали участие в протестах с изменением электорального поведения.

В некотором роде обратное исследование провели итальянские политологи Е. Паван и А. Маинарди [Pavan, Mainardi, 2018]. Они изучили мобилизацию итальянского движения против гендерного насилия Non Una Di Meno (NUDM). Прежде всего они собрали сет твитов, созданных в ходе двух национальных акций протеста, организованных движением. Затем исследователи реконструировали два набора сетей, социальную и семантическую, созданные участниками движения вокруг национальной забастовки 8 марта 2017 г. и б) вокруг национального марша, организованного в Риме 25 ноября 2017 г. Опираясь в основном на инструменты сетевого анализа, Паван и Маинарди пришли к выводу, что структурные особенности сетевых структур оказались устойчивыми: несмотря на расширение изменений протестного репертуара движения, сетевые структуры, возникшие в результате внедрения цифровых медиа, оставались немногочисленными и сильно кластеризованными. Кроме того, несмотря на слабость и малочисленность этих структур, онлайновое общение, развернувшееся вокруг ноябрьского марша, после нескольких месяцев активности NUDM по консолидации своего коллективного проекта по борьбе с гендерным насилием стало более инклюзивным, поскольку лишь меньшинство участников оставалось изолированным или вело отдельные разговоры. Приобретя более инклюзивные черты, онлайновое общение вокруг NUDM также стало более заметно ориентироваться на официальную страницу движения.

## Критика использования Big data в политических исследованиях

Несмотря на наличие определенных позитивных результатов в применении больших данных многие исследователи относятся к ней критично по целому ряду оснований. Традиционная критика использования больших данных в социальных и политических исследованиях построена на том, что данные, которые нам представляют, имеют как минимум «неидеальный характер». Часто исследователи видят паттерны там, где их в действительности нет просто потому, что в гигантском объеме данных можно что-нибудь найти, что соединяется со всем во всех направлениях [Boyd, 2012, p. 668]. В ходе процесса сбора данных самими платформами (например, социальных сетей) данные обрабатываются в несколько шагов (аккумуляция, очистка, сбор в базу для дальнейшей обработки) и на каждом этапе возможны ошибки и искажения (как произвольные, так и случайные) [Wagschal, 2020, p. 276].

Другой проблемой является то, что зачастую исследователям доступны только API (Application Programming Interface – программный интерфейс приложения). Говоря упрощенно, API – это контракт, который предоставляет программа, сайт, платформа о том, как с ней обращаться и что она может дать[1]. Поскольку доступны только специально отобранные и подготовленные владельцами сайта API данные, возможности проведения исследований оказываются ограничены. Американский политолог К. Мангер напоминает, что Facebook, Twitter постоянно меняются (в смысле алгоритмов обработки и выдачи данных)[2]. Изменения в работе платформы ведут к изменению поведения. В свою очередь это делает лонгитюдные исследования на основании данных этих сетей слабо валидными.

Значительная проблема обнаружилась даже в данных, собираемых уважаемыми агрегаторами онлайн – баз данных. Проекты по сбору данных, вроде Militarized Interstate Dispute (MID), Uppsala Conflict Data Program (UCDP), Armed Conflict Location and Event Dataset (ACLED) (и Mass Mobilization in Autocracies

[1] См. также: Что такое API? – Режим доступа: https://habr.com/ru/articles/464261/ (дата посещения: 01.07.2023).

[2] См. также: Big data in Political science. – Режим доступа: https://www.youtube.com/watch?v=gdctyW6ghgg&t=4443s (дата посещения: 01.07.2023).

Database (MMAD), предоставляют свои услуги исследователям, снижая входной барьер для количественного анализа. Как показал анализ, выполненный американскими политологами [Karstens, Soules, Dietrich, 2023], качество и полнота баз данных оставляют желать лучшего. Источники, доступные через них, меняются с течением времени. Периодически и без предупреждения пользователей истекает срок действия контрактов или происходит их перезаключение, в результате чего некоторые источники исчезают из базы данных.

Кроме того, оказалось, что наличие одного и того же источника в двух базах данных не гарантирует доступ к одним и тем же публикациям. Например, Nexis Uni и Factiva оба имеют Синьхуа в своих списках источников, но они содержат разные публикации данного издания. Таким образом, результаты поиска могут варьироваться в зависимости от скрытых от исследователя факторов. И хотя крупные события вряд ли будут упущены из-за их значительного освещения, более мелкие события, которые привлекают меньше внимания СМИ, скорее всего, будут оставаться вне поля зрения ученых.

Использование метаданных мобильных устройств само по себе не может дать исчерпывающего представления о том, кто и почему присоединяется к коллективным акциям протеста [Rotman, 2022]. Сценарий, в котором власти, как подозревают протестующие, могут использовать данные о местоположении для их идентификации и наказания, может заставить участников демонстрации оставлять свои мобильные телефоны дома или выключать их во время демонстраций. Другие ограничения зависят от особенностей сбора и распространения данных о местоположении отдельными сотовыми сетями и поставщиками данных, а также от ограничений, накладываемых регулирующими органами.

Отдельный вопрос возникает к качеству и даже необходимости электоральных предсказаний. К примеру, в США, стране, которая позиционирует себя как «светоч демократии», на выборах в Палату представителей в 2016 г. только 35 мест из 435 были получены хоть в какой-то конкурентной борьбе. При этом лишь 17 из них были избраны с перевесом до 5%, а еще 18 – с перевесом до 10% [Чизмен, Клаас, 2021, с. 65–73]. В таком случае предсказание результатов выборов на основе данных социальных сетей с точностью в 90% не является большим достижением.

Есть целый ряд возражений с чисто технической стороны организации работы исследовательских алгоритмов. Например, кластеризация методом k-средних плоха тем, что каждый элемент данных может быть связан только с одним кластером, тогда как элемент часто может находиться между двумя центрами, что делает его включение в один из них равновероятным. Другим допущением кластерного анализа является сферическая форма кластера. В случае, если кластер имеет иную форму, он будет автоматически усечен, а усеченные элементы попадут в другой кластер. Метод k-средних не допускает пересечения кластеров или их вложения друг в друга [Ын, Су, 2022, с. 48]. Метод опорных векторов классифицирует элементы, исходя из того, с какой стороны границы дифференциации они оказались. В ситуации, когда элементы данных сильно перекрываются обеими группами, то те из них, что ближе к границе, могут быть классифицированы ошибочно. Кроме того, метод не дает информацию о вероятности ошибочной классификации каждого отдельного элемента [Ын, Су, 2022, с. 125]. Недостаток деревьев решений вытекает из их достоинства: они слишком восприимчивы к переобучению. Градиентный бустинг позволяет уточнять прогнозы, однако он зачастую очень сложен для визуализации [Ын, Су, 2022, с. 135].

В литературе есть крайне серьезное возражение к использованию больших данных для прогнозирования выборов со стороны статистической науки. Так, в исследовании, сравнивающем точность предсказаний обычного опроса общественного мнения с 60% ответов и сет данных, охватывающий 80% населения, профессор статистики Гарвардского университета показал, что первый набор данных будет точнее [Meng, 2018]. В статье подчеркивается важность понимания закона больших совокупностей (law of large populations,), который гласит, что при увеличении объема выборки среднее значение выборки дает более надежную оценку среднего значения совокупности. Однако автор утверждает, что в сфере больших данных этот закон может быть обманчивым из-за предвзятости отбора. Парадокс больших данных, по мнению автора, связан с тем, что, хотя они предоставляют огромное количество информации, в них может отсутствовать необходимый контроль и случайность, требуемые для корректного статистического вывода.

На более абстрактном уровне возникает ряд проблем общетеоретического плана. Прежде всего речь идет о парадоксе

Р. Лапьера, который показывает автономию установок и поведения. По данным многочисленных этнометодологических исследований, символы, нарративы, дискурсы, верования, коды, репрезентации, планы существуют исключительно в мире слов и ограниченно влияют на реальные практики [Вахштайн, 2021]. Так, несмотря на то, что в Сети пользователь может демонстрировать приверженность одним ценностям, регулярно ставить лайки, например, ультралиберальной позиции, в офлайн-режиме его политическое действие может принять совсем другой, консервативный характер.

С этим парадоксом тесно связано и другое общетеоретическое социологическое возражение, которое исходит из теории действия. В отличие от старых «интеллектуалистских» теорий (например, теория действия Т. Парсонса), которые требуют обращения к привычному «психологическому» аппарату причин, намерений, убеждений и желаний для объяснения действий в новых диспозициональных объяснениях, основанных на привычках, действия вместо этого объясняются путем ссылки на тенденцию или склонность к повторению организованных паттернов действий в конкретных условиях надежным (но не детерминированным) образом, учитывая историю агента [Lizardo, 2021]. Это относится не только к действиям в физическом, офлайновом мире, но и к ментальным процессам или онлайн-активностям. Сутью человека фактически выступает набор его привычек. Из этого можно сделать вывод, что агрегированные данные могут описывать набор привычек, характерных только для той среды, в которых они и собираются, но, мало соотносится с другими сферам жизни существования, которые исследователи пытаются раскрыть.

## Вместо заключения. Перспективы использования больших данных в политических исследованиях

Данный краткий обзор современного использования больших данных и машинного обучения показывает, что они суть не волшебная кнопка, по нажатию которой политолог получает автоматический результат о настоящем, прошлом и будущем политической жизни. Напротив, это большое множество методов анализа и валидации результатов, имеющих свое ограниченное применение.

В ближайшие годы объемы данных и утонченность алгоритмов их анализа продолжат расти. Тем не менее, вне всякого сомнения, будут выходить «ленивые» исследования, в которых авторы попытаются каким-то образом наложить разные базы данных друг на друга и получить некую неожиданную корреляцию, выдав ее за научный прогресс. Иллюзия дешевизны, быстроты и мнимой общедоступности больших данных также продолжит приводить к появлению «одномерных публикаций», в которых на плохих данных с помощью одного самого просто алгоритма (причем взятого с чужого кода без собственной рефлексии) будут делаться далеко идущие выводы о политических процессах.

Если отбросить пессимизм в сторону, то по публикациям в ведущих методологических журналах, вроде Political analysis, проводимым политологическими ассоциациями мероприятиям можно выделить тенденцию большей осторожности и методологической точности в применении как больших данных, так и машинного обучения у тех, кто находится на передовом крае исследований.

Весьма вероятно, что в среднесрочной перспективе ответственные ученые будут все больше сосредоточиваться на использовании компьютационных методов для анализа цифровых сред, где, собственно, эти данные и генерируются. Будет расти доля публикаций об онлайн-политическом участии. Можно предположить, что будет крепнуть движение верификации данных и репликации исследований. Ожидается и появление большего количества публикаций, проясняющих онтологические и эпистемологические характеристики нового «цифрового» знания.

Другим важным направлением станет применение методов машинного обучения для вспомогательных задач сортировки, например, архивных данных, кластеризации, тематизации, что, вероятно, внесет существенный вклад в исследования в духе исторического институционализма или для применения process tracing в целях анализа прошедших событий.

Изощренность новых методов анализа, рост требований к публикациям в высокорейтинговых журналах ставят вопросы о будущем содержании учебных программ и квалификации политологов. Данных, методов и стоящих за ними математики, статистики становится чересчур много, тогда как количество часов зачастую остается тем же или даже сокращается. К тому же возникают вопросы, как быть с преподаванием других субдисциплин, кото-

рые считались основой канона политического знания. Уже в ближайшее время встанет вопрос об интеграции ChatGPT не только для решения вышеупомянутых примитивных задач, но и возможности его использования для подготовки более серьезных исследований.

**I.E. Kochedykov***

**On the experience of applying Big data in political science**

*Abstract*. The review article is devoted to the analysis of successes and challenges of Big Data application in political science. The first part discusses the ontological and epistemological foundations of Big Data and machine learning application in political science. In the second part, the author reviews representative results of political science research using Big Data. The third part deals with criticism and limitations of Big Data in political research. The author shows besides purely technical problems, such as incompleteness of available data, distortions due to the presence of bots, there are sufficient limitations to the application of big data for analyzing dispositional political actions.

*Keywords:* Big data; algorithms; machine learning; ontology and epistemology of big data; theory of action.

*For citation:* Kochedykov I.E. On Big data application for political science research. *Political science (RU).* 2023, N 4, P. 226–251. DOI: http://www.doi.org/10.31249/poln/2023.04.09

## References

Anderson C. The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine.* 2008, July 16, P. 1–2. Mode of access: http://statlit.org/pdf/2008EndOfTheory-DataDelugeMakesScientificMethodObsolete-WiredMagazine.pdf (accessed: 10.07.2023).

Akhremenko A., Petrov A. Anger, identity or efficacy belief? Dynamics of motivation and participation in 2020 Belarusian protests. *Polis. Political Studies*, 2023, N 2, P. 138–153. (In Russ.) DOI: https://doi.org/10.17976/jpps/2023.02.10

Ash E., Gauthier G., Widmer P. Relatio: Text semantics capture political and economic narratives. *Political Analysis*. 2023, First View. P. 1–18. DOI: https://doi.org/10.1017/pan.2023.8

Athey S. Beyond prediction: Using big data for policy problems. *Science*. 2017. N 355, P. 483–485. DOI: https://doi.org/10.1126/science.aal432

* **Kochedykov Ivan**, HSE University (Moscow, Russia), e-mail: Kochedykov.ivan@gmail.com

Barclay F., Pichandy C., Venkat A., Sudhakaran S. India 2014: Facebook 'like' as a predictor of election outcomes. *Asian Journal of Political Science*. 2015, Vol 23, N 2, P. 134–160. DOI: https://doi.org/10.1080/02185377.2015.1020319

Benoit K., Munger K., Spirling A. Measuring and explaining political sophistication through textual complexity. *American Journal of Political Science*. 2019, Vol. 63, N 2, P. 491–508. DOI: https://doi.org/10.1111/ajps.12423

Bond R., Messing S. Quantifying social media's political space: Estimating ideology from publicly revealed preferences on Facebook. *American Political Science Review*. 2015, Vol. 109, N 1, P. 62–78. DOI: https://doi.org/10.1017/S0003055414000525

Bonica A. A data-driven voter guide for US elections: Adapting quantitative measures of the preferences and priorities of political elites to help voters learn about candidates. *RSF: The Russell Sage Foundation Journal of the Social Sciences*. 2016, Vol. 2, N 7, P. 11–32. DOI: https://doi.org/10.7758/rsf.2016.2.7.02

Boullier D. The social sciences and the traces of big data. *Revue francaise de science politique*. 2015, Vol. 65, N 5, P. 805–828. DOI: https://doi.org/10.48550/arXiv.1607.05034

Boyd D., Crawford K. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*. 2012, Vol. 15, N 5, P. 662–679. DOI: https://doi.org/10.1080/1369118X.2012.678878

Ceron A., Curini L., Iacus S. Using sentiment analysis to monitor electoral campaigns: Method matters – evidence from the United States and Italy. *Social Science Computer Review*. 2015, Vol. 33, N 1, P. 3–20. DOI: https://doi.org/10.1177/0894439314521983

Cheeseman N., Klaas B. *How to rig and election*. Moscow: Bombora, 2021, 320 p. (In Russ.).

Ceron A., Curini L., Iacus S. Social media and elections: A meta-analysis of online-based electoral forecasts. *In:* Arzheimer K., Evans J., Lewis-Beck M.S. (eds.) *Sage Handbook of electoral behaviour*. London: SAGE Publications Ltd, 2017, P. 883–903. DOI: https://doi.org/10.4135/9781473957978

Chatsiou K., Mikhaylov S.J. Deep learning for political science. *arXiv preprint arXiv:2005.06540*. 2020. Mode of access: https://arxiv.org/abs/2005.06540 (accessed: 10.07.2023).

Colaresi M., Mahmood Z. Do the robot: Lessons from machine learning to improve conflict forecasting. *Journal of Peace Research*. 2017, Vol. 54, N 2, P. 193–214. DOI: https://doi.org/10.1177/0022343316682065

Conover M.D., Davis C., Ferrara E., McKelvey K., Menczer F., Flammini A. The geospatial characteristics of a social movement communication network. *PloS one*. 2013, Vol. 8, N 3. DOI: https://doi.org/10.1371/journal.pone.0055957

Gandomi A., Haider M. Beyond the hype: big data concepts, methods, and analytics. *International journal of information management*. 2015, Vol. 35, N 2, P. 137–144. DOI: https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Grimmer J., Stewart B.M. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*. 2013, Vol. 21 N. 3, P. 267–297. DOI: doi:10.1093/pan/mps028

Grossman J., Pedahzur A. Political science and big data: Structured data, unstructured data, and how to use them. *Political science quarterly*. 2020, Vol. 135, N 2, P. 225–257. DOI: https://doi.org/10.1002/polq.13032

Karstens M., Soules M.J., Dietrich N. On the Replicability of Data Collection Using Online News Databases. *PS: Political Science & Politics*. 2023, Vol. 56, N 2, P. 265–272. DOI:10.1017/S1049096522001317

Katagiri A., Min E. The credibility of public and private signals: A document-based approach. *American Political Science Review*. 2019, Vol. 113, N 1, P. 156–172. DOI: https://doi.org/10.1017/S0003055418000643

Sang E.T., Bos J. Predicting the 2011 Dutch senate election results with twitter. *Proceedings of the workshop on semantic analysis in social media*. 2012. – Mode of access: https://www.let.rug.nl/bos/pubs/TjongBos2012EACL.pdf (accessed: 11.06.2023).

King G., Pan J., Roberts M.E. How censorship in China allows government criticism but silences collective expression. *American political science Review*. 2013, Vol. 107, N 2, P. 326–343. DOI: https://doi.org/10.1017/S0003055413000014

Kitchin R., McArdle G. What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*. 2016, Vol. 3, N 1. DOI: https://doi.org/10.1177/205395171663113

Lazer D., Radford J. Data ex machina: introduction to big data. *Annual Review of Sociology*. 2017, Vol. 43, P. 19–39. DOI: https://doi.org/10.1146/annurev-soc-060116-053457

Lizardo O. Habit and the Explanation of Action. *Journal for the Theory of Social Behaviour*. 2021, Vol. 51, N 3, P. 391–411. DOI: https://doi.org/10.1111/jtsb.12273

Meng X-L. Statistical paradises and paradoxes in big data (i) law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics*. 2018, Vol. 12, N 2, P. 685–726. DOI: https://doi.org/10.1214/18-AOAS1161SF

Montgomery J.M., Olivella S., Potter J.D., Crisp B.F. An informed forensics approach to detecting vote irregularities. *Political Analysis*. 2015, Vol. 23, N 4, P. 488–505. DOI: https://doi.org/10.1093/pan/mpv023

Mueller H., Rauh C. Reading between the lines: Prediction of political violence using newspaper text. *American Political Science Review*. 2018, Vol. 112, N 2, P. 358–375. DOI: https://doi.org/10.1017/S0003055417000570

Ng A., Soo K. *Numsense! Data science for the layman: no math added*. Piter, 2022, 208 p. (In Russ.).

Park B., Greene K., Colaresi M. Human rights are (increasingly) plural: Learning the changing taxonomy of human rights from large-scale text reveals information effects. *American Political Science Review*. 2020, Vol. 14, N 3, P. 888–910. DOI: https://doi.org/10.1017/S0003055420000258

Pavan E., Mainardi A. Striking, Marching, Tweeting. Studying how online networks change together with movements. *Partecipazione e conflitto*. 2018, Vol. 11, N 2, P. 394–422. DOI: 10.1285/i20356609 v11 i2 p394

Pietsch W. Aspects of theory-ladenness in data-intensive science. *Philosophy of Science*. 2015, Vol. 82, N 5, P. 905–916. DOI: https://doi.org/10.1086/683328

Pietsch W. *On the epistemology of data science: Conceptual tools for a new inductivism*. Berlin; Heidelberg: Springer-Verlag. 2021, Vol. 148, 295 p. DOI: https://doi.org/10.1007/978-3-030-86442-2

Reis, J., Correia A., Murai F., Veloso ABenevenuto F. Supervised learning for fake news detection. *IEEE Intelligent Systems*. 2019, Vol. 34, N 2, P. 76–81. DOI: https://doi.org/10.1109/MIS.2019.2899143

Rettberg J.W. Algorithmic failure as a humanities methodology: Machine learning's mispredictions identify rich cases for qualitative analysis. *Big Data & Society*. 2022, Vol. 9, N 2. DOI: https://doi.org/10.1177/20539517221131290

Rotman A., Shalev M. Using location data from mobile phones to study participation in mass protests. *Sociological Methods & Research*. 2022, Vol. 51, N 3, P. 1357–1412. DOI: https://doi.org/10.1177/0049124120914926

Salganik M.J. *Bit by bit: Social research in the digital age*. Princeton: Princeton University Press, 2019, 448 p.

Vakhshtayn, V. *Technics, or the Charm of Progress.* St. Petersburg: European University press, 2021, 156 p. (In Russ.).

Wagschal U. Ettensperger F. Big Data in Social Sciences. *In:* Berg-Schlosser D., Badie B., Morlino L. (eds.) *The SAGE Handbook of Political Science.* London: SAGE Publications Ltd, 2020, P. 272–287. DOI: 10.4135/9781529714333. n19

## Литература на русском языке

*Ахременко А.С., Петров А.П.* Гнев, идентичность или вера в успех? Динамика мотивации и участия в белорусских протестах 2020 года // Полис. Политические исследования. – 2023. – № 2. – С. 138–153. DOI: https://doi.org/10.17976/jpps/2023.02.10

*Вахштайн В.С.*[1] Техника, или Обаяние прогресса. – СПб: Изд-во Европейского университета в Санкт-Петербурге, 2021. – 156 с.

*Чизмен Н., Клаас Б.* Как почти честно выиграть выборы. – М.: Бомбора: 2021, 320 с.

*Ын А., Су К.* Теоретический минимум по Big Data. Всё, что нужно знать о больших данных. – СПб.: Питер, 2022. – 208 с.

[1] Внесен в список иностранных агентов.