ЦИФРОВЫЕ ТЕХНОЛОГИИ В СИСТЕМЕ СОЦИАЛЬНО-ГУМАНИТАРНЫХ НАУК: ОПЫТ УЗБЕКИСТАНА, ЗАРУБЕЖНЫХ СТРАН И СТРАН ЦЕНТРАЛЬНОЙ АЗИИ

НОВЫЙ ЭТАП ЦИФРОВОЙ ТРАНСФОРМАЦИИ ИСТОРИЧЕСКИХ ИССЛЕДОВАНИЙ: ВОЗМОЖНОСТИ ТЕХНОЛОГИЙ БОЛЬШИХ ДАННЫХ И ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Бородкин Л.И.
д.и.н., член-корр. РАН, профессор,
МГУ имени М.В. Ломоносова

Процесс цифровой трансформации исторических исследований и образования в последние годы приобретает новые черты. В то время как традиция использования историками статистических методов и компьютерных технологий насчитывает уже шесть десятилетий, с началом XXI века пришло время заметного обновления инструментария, пригодного доля обработки и анализа данных исторических источников. Речь идет о стремительном развитии науки о данных, появлении новых «интеллектуальных» методов и технологий работы с данными (когда данные стали «большими»).

Наука о данных (Data Science). Этот термин получил широкое распространение с начала 2000-х гг., когда привычные методы обработки и анализа данных, основанные на методах математической статистики, стали активно дополняться методами и технологиями искусственного интеллекта, а также «больших данных». Сегодня эта область рассматривается как часть современной информатики, ориентированной на работу с данными, представленными в цифровой форме, неотъемлемой ее частью стали методы и технологии искусственного интеллекта (ИИ), включая машинное обучение, искусственные нейросети и большие данные; при этом обработка и анализ данных требуют больше компетенций в информационных технологиях, программировании, методах визуализации. И хотя большинство таких специалистов работают в сфере бизнес-приложений, спрос на них растет и в социально-

гуманитарных исследованиях. Сегодня уже можно говорить о тенденции введения курса «Наука о данных» в учебный план образовательных программ гуманитарного профиля. Так, на истфаке МГУ с 2021 г. введен обязательный для всех магистрантов истфака курс «Наука о данных и искусственный интеллект в исторических исследованиях», а специализация на кафедре исторической информатики включает такие дисциплины как «Моделирование исторических процессов» и программирование на языках **R** и Python.

Рассмотрим кратко опыт использования историками новых методов науки о данных и технологий ИИ.

Большие данные в исторических исследованиях

Имеют ли историки опыт применения методов искусственного интеллекта? Да, такой опыт имеется, и он, в общем-то, немалый, берущий начало еще в прошлом веке. Первые публикации в этой области появились в 1980-х – 1990-х гг. и были связаны с применением когнитивных компьютерных моделей понимания текста, а также с разработкой экспертных систем [2].

Прежде чем обратиться к характеристике современного этапа использования концепций и технологий искусственного интеллекта (ИИ) в работах историков, рассмотрим кратко понятие о *больших данных,* существенное для понимания структуры исследований XXI века в области искусственного интеллекта. Является ли проблема больших данных актуальной для историков?

Если говорить об использовании оцифрованных исторических данных, то и в России, и в других странах абсолютное большинство возникающих задач можно решить на «продвинутых» персональных компьютерах (или, в особых случаях, на суперкомпьютерах). В то же время появляются задачи, для решения которых мощностей такой вычислительной техники уже не хватает. Это порождает вопрос: работают ли историки сегодня с большими данными (*Big Data*)? Ответ зависит от того, какого определения мы придерживаемся. Недавно в России был принят Национальный стандарт РФ «Большие данные», в котором подробно определены соответствующие дефиниции и технологии. *Big Data* определяются как большие массивы данных, с учетом таких их характеристик, как объем, разнообразие, скорость обработки и/или вариативность, которые требуют использования технологии масштабирования для эффективного хранения, обработки, управления и анализа. Масштабирование подразумевает возможность расширения репозиториев поступающих *потоковых* данных и их обработку на параллельно работающих распределенных ресурсах.

Эти технологии альтернативны традиционным технологиям баз данных. Хотя строгое определение больших данных подразумевает их потоковую природу (непрерывное их поступление), в практике социально-гуманитарных исследований такие данные встречаются редко (исключение сводится, пожалуй, к исследованиям социальных сетей с их непрерывным информационным потоком, порождаемым Интернет-сервисами); зато в исторических исследованиях возникает необходимость анализа огромного объема разнородных источниковых данных, работа с которыми не может производиться с использованием компьютеров и стандартного программного обеспечения, привычных для пользователей-историков. Даже при отсутствии потоковых данных хранение и обработка информационного массива, объем которого исчисляется десятками и сотнями терабайт, требует специальных технических и программно-алгоритмических решений, поскольку стандартные аппаратные и программные средства в подобных ситуациях бесполезны. При этом данные должны быть распределены, как правило, по вычислительным узлам, кластерам компьютеров.

В широком смысле «Большие данные» рассматривают как социально-экономический феномен, связанный с появлением технологических возможностей анализировать огромные массивы данных. В качестве определяющих характеристик для больших данных традиционно выделяют «три V»: объём (англ. volume, в смысле величины физического объёма данных), скорость (velocity в смыслах как скорости прироста, так и необходимости высокоскоростной обработки и получения результатов), многообразие (variety, в смысле возможности одновременной обработки различных типов структурированных и неструктурированных данных). Встречаются в публикациях предложения учитывать и пять V, и семь V.

Материалы ряда недавних конференций по технологиям Big Data позволяют сделать вывод о том, что сегодня в отмеченных ситуациях говорят о допустимом расширении определения больших данных, когда данные могут не иметь потоковой природы, но объем их настолько велик, что они не могут храниться и обрабатываться стандартными средствами. Характерной особенностью таких данных в *исторических* исследованиях является не только огромный объем источниковых данных, лежащих в основе масштабных компаративных исследовательских проектов, реализуемых историками в составе международных коллабораций, но и вариативность, поливидовой характер используемых источниковых комплексов.

Исходя из приведенной трактовки больших данных, на сегодняшний день можно обратиться к нескольким масштабным историко-ориентированным проектам, в которых используются большие данные (в указанном выше понимании).

Международный проект "Mining Microdata: economic opportunity and spatial mobility in Britain, Canada and The United States, 1850–1911". Этот проект, завершенный в 2014 г., базировался на огромном информационном ресурсе, использовавшем первичные данные переписей населения США, Канады и Британии 1850-1911 гг. (более 100 млн записей). Проведенные в рамках проекта исследования внесли существенный вклад в длительные дебаты историков об эволюции социальной структуры населения в Северной Америке и Великобритании. В ходе работ по проекту были предложены также оригинальные компьютеризованные методики источниковедческого анализа больших массивов данных переписей (с использованием алгоритмов *record linkage*, включающих методы искусственного интеллекта) [4].

Проект CLARIAH (Common Lab Research Infrastructure for the Arts and Humanities). Этот проект вносит существенный вклад в создание современной инфраструктуры гуманитарных и социальных наук на базе Международного института социальной истории (IISH) в Амстердаме CLARIAH (// https://www.clariah.nl/en/about/about-clariah).

Реализация проекта проводится при поддержке голландского научного фонда (NWO) и предусматривает два этапа: 2014-2018 гг. и 2019-2024 гг.; объем финансирования на каждом этапе – 12 млн. евро. CLARIAH рассматривается как голландская часть учрежденного в 2014 г. Европейского сетевого консорциума DARIAH (Digital Research Infrastructure for the Arts and the Humanities), целью которого является совершенствование и развитие европейской цифровой инфраструктуры для исследований в области гуманитарных наук DARIAH-EU (// https://www.dariah.eu/about/dariah-in-nutshell/).

Проект «Венецианская машина времени» (*The Venice Time Machine*). Этот масштабный проект, основанный на цифровизации огромного архивного материала, ориентирован на дальнейшее многоцелевое комплексное использование оцифрованных данных, был инициирован в 2012 г., а активная работа началась в 2014 г. в рамках коллаборации Государственного архива Венеции, Швейцарского федерального технологического института в Лозанне (EPFL) и Университета Ка´ Фоскари в Венеции. Руководитель проекта, проф. Фредерик Каплан, IT-специалист, ученый из EPFL, отметил, что документы, подлежащие оцифровке, занимают более 80 км полок в Государственном архиве Венеции (подробнее об этом см. статью в журнале *Nature* [5]). Общей целью проекта является создание многомерной динамической «модели» Венеции путем создания открытого цифрового архива культурного наследия города, охватывающего более 1000 лет его эволюции (Venice Time M55achine - EPFL). Методология проекта предполагает, что существует связь между людьми, чьи имена указаны в одном и том же

документе, и это позволяет выявлять социальную сеть каждого человека, «вписывая» ее в мега-сеть большого исторического социума.

«Венецианская машина времени» дала старт еще более амбициозному европейскому проекту по созданию и объединению подобных «машин времени» в целом ряде исторических центров культуры и торговли Европы (Time Machine Europe) (включая Париж (1000 – 2000 гг.), Дрезден (1200 – 2000 гг.), Будапешт (1660 – 1990 гг.), Иерусалим (2000 до н.э. – 2000 н.э) и др. Такой масштаб интегрального европейского проекта «машин времени» (с бюджетом 1 млрд евро (!)) потребовал формирования координационного Совета в рамках “TimeMachine Organisation”; президентом организации является Ф. Каплан, а основное финансирование мега-проекта, имеющего целью «создание планетарного масштаба инфраструктуры, обеспечивающей прямой и свободный доступ к Большим данным о Прошлом», осуществляется Europeana Foundation.

ИИ в исторических исследованиях

Наибольший интерес в этой области вызывает применение методов машинного обучения в его наиболее распространенном варианте, связанном с возможностями искусственных нейронных сетей. В XXI в. появились исследовательские проекты, в которых машинное обучение использовалось (уже на новом технологическом витке цифровых технологий и алгоритмов) для решения различных задач в области археологии, изучения и сохранения культурного наследия, в источниковедческих и конкретно-исторических исследованиях. Библиография таких работ к середине 2022 г. насчитывает уже сотни публикаций. Рассмотрим лишь некоторые из них, дающих представление о возможностях применения методов и технологий ИИ историками.

1. Актуальной задачей является *распознавание* отсканированных рукописных и старопечатных текстов – ведь ежедневно в мире делаются доступными пользователю тысячи страниц оцифрованных, но не являющихся машиночитаемыми источников, даже информационный поиск в таких случаях невозможен. Надо признать, что существующие методы распознавания знаков таких текстов пока еще не обладают достаточной степенью универсальности. Машинное обучение ориентируется, как правило, на распознавание текстов, написанных на определенном языке, в определенную эпоху, в той или иной форме письма.

В этом контексте представляет интерес научно-исследовательский проект «Автографы Петра Великого: Чтение технологиями искусственного интеллекта», инициированный в 2020 г. Российским историческим обществом и ПАО «Сбербанк». В работе над проектом приняла активное участие исследовательская группа Санкт-Петербургского института

истории РАН. В основу источниковой базы были положены рукописи из собраний института и РГАДА. Эта база была сформирована из материалов последних томов «Писем и бумаг Петра Великого», охватывавших период с 1709 по 1713 год. Ход работы по проекту описан в публикации членов коллектива участников [1]. Существенным итогом первого этапа работы стало размещение компьютерной программы по распознаванию петровских текстов на сайте «Digital Петр» (https://www.sber.ru/digital-petr/) [7]. На этом сайте пользователь может за несколько секунд получить в удобном и понятном формате расшифровку рукописей Петра Первого. Для тех, кто хочет понять принцип работы алгоритма, на сайте предусмотрен режим ознакомления и описание того, как работает нейросеть. Как отмечают разработчики сайта, посетители могут загрузить рукопись Петра I самостоятельно, а могут воспользоваться функцией расшифровки заранее заготовленных рукописей.

2. Атрибуция древних текстов (эпиграфика). Методы искусственного интеллекта в последние годы нашли применение и в такой традиционной области изучения античной истории, как эпиграфика. В течение прошедших столетий многие надписи были повреждены до неразборчивости, перенесены далеко от своего первоначального места, а дата их написания является предметом дискуссий специалистов. Новый импульс в исследованиях таких проблем дают технологии машинного обучения. Пример такого исследования представлен в совместной работе ученых из Великобритании, Греции и Италии – историков античности и IT-специалистов, опубликованной в журнале *Nature* [6].

Предложенная система *Ithaca* рассматривается ее разработчиками в качестве образца для современного подхода к решению задач эпиграфики. Система основана на архитектуре нейронной сети, которая использует механизм взвешивания влияния различных частей ввода (таких как символы, слова) на принятие решений моделью.

3. *Классификация занятий репрессированных в 1930-е – 1940-е годы.* Рассмотрим конкретно-историческое исследование, на примере которого можно выявить проблемы, которые историк должен решать при использовании технологий ИИ. В работе выпускницы кафедры исторической информатики МГУ Л.А. Лягушкиной [3] изучается вопрос о классификации занятий репрессированных в СССР, сведения о которых внесены в базу данных «Жертвы политического террора в СССР», включающей более трех миллионов записей о людях, которые подвергались репрессиям в советское время. На первом этапе исследования автор располагала данными о 65 тыс. человек, репрессированных в ходе «Большого террора» (1937-1938) в пяти регионах РСФСР, с указанием их занятий и категорий их социального положения. На следующем этапе (для изучения гендерного аспекта репрессий) автор расширила временной интервал исследования (включив 1941-1945 гг.) и

географический масштаб работы (добавив еще 20 регионов), что привело к необходимости классифицировать занятия еще примерно 320 тыс. человек. Применение алгоритма обучения нейросети позволило разметить 350 тыс. новых записей из базы данных. Результаты работы модели (на тестовой выборке) показали, что в целом модель довольно успешно «предсказала» классы записей по социальной группе репрессированных: в 95% случаев классы на тестовой выборке были определены правильно.

* * *

Быстрое развитие концепций «Больших данных», их адаптация с учетом специфики различных наук и массовая цифровизация данных во всех областях научного знания оказали определенное воздействие и на гуманитарные науки, включая историю. Методы ИИ, основанные на использовании машинного обучения, искусственных нейросетей, требуют, как правило, обучающих выборок большого объема, что создает спрос на формирование больших массивов данных.

В течение последнего десятилетия реализован ряд крупных междисциплинарных исследовательских проектов историков, в рамках которых сформированы большие массивы (базы) данных, на основе которых решаются взаимосвязанные задачи, позволяющие получить новые ответы на «старые» вопросы и поставить новые вопросы, исходя из диалектического принципа «количество переходит в качество». Как показывает опыт, реализация таких проектов (нечастых пока) требует масштабного финансирования и четкой логистики в реализации поставленных задач.

Использование методов ИИ в исторических исследованиях требует совместной работы историков и IT-специалистов. Не следует полагать, что обученная машина сделает всё сама, и нейросеть даст историку «правильный» результат. В такой работе есть немало «подводных камней», которые можно учесть в междисциплинарном взаимодействии историков и IT-специалистов.

Литература:

1. Базарова Т., Димитров Д., Потанин М., Проскурякова М. Распознать и транскрибировать: Автографы Петра Великого и технологии искусственного интеллекта // Вестник «Воронцово поле». 2020. № 4. С. 64–71

2. Бородкин Л.И. Методы искусственного интеллекта: новые горизонты исторического познания // Информационный Бюллетень Комиссии по применению математических методов и ЭВМ в исторических исследованиях при отделении истории РАН. 1992. № 5. С. 4-16.

3. Лягушкина Л.А. Опыт классификации социального положения репрессированных в СССР с помощью метода опорных векторов // Историческая информатика. 2022. № 1. – С. 128 - 139.

4. Шурер К., Пенькова Т. Методы стандартизации и классификации записей о месте рождения и профессиональной деятельности в данных переписи Великобритании 1851–1911 гг. // Историческая информатика. – 2016. – №3-4. – С.41-56.

5. Abbott, A. The ‘Time Machine’ reconstructing Ancient Venice’s social networks. Nature 546, 341–344 (2017)

6. Assael, Y., Sommerschield, T., Shillingford, B. et al. Restoring and attributing ancient texts using deep neural networks. Nature 603, 280–283 (2022).

СОЗДАНИЕ ЭЛЕКТРОННЫХ КОПИЙ АРХИВНЫХ ДОКУМЕНТОВ – СОВРЕМЕННЫЙ МЕТОД СОХРАНЕНИЯ ИСТОРИКО-КУЛЬТУРНОГО НАСЛЕДИЯ И РАСШИРЕНИЯ ДОСТУПА К НЕМУ

Юмашева Ю.Ю.
Член Комиссии по научно-методической работе
Федерального архивного агентства,
Москва, Россия

Одна из центральных задач, которая стоит перед архивами, библиотеками и музеями всего мира это создание электронных фондов пользования копий архивных документов, книг и музейных предметов.

Для архивов Российской Федерации эта задача была сформулирована в государственной программе «Информационное общество» в 2010 г. и подтверждена в «Программе цифровой трансформации», которая утверждена в 2020 гг.

Федеральное архивное агентство активно осуществляет работы по созданию электронных копий архивных документов. При этом оцифровкой занимаются сами архивы, а также компании-подрядчики.

В рамках решения этой задачи в 2012 г. по заказу Федерального архивного агентства был разработан комплекс нормативно-методических актов, регулирующих процессы оцифровки архивной документации.

В состав данного комплекса входили несколько документов. В том числе:

- Методические рекомендации по созданию электронных копий документов АФ РФ на бумажной основе и управлению полученным