НАУКА О КУЛЬТУРЕ? ИЗУЧЕНИЕ БОЛЬШИХ КУЛЬТУРНЫХ ДАННЫХ: СОЦИАЛЬНАЯ ИНФОРМАТИКА, ЦИФРОВАЯ ГУМАНИТАРИСТИКА И КУЛЬТУРНАЯ АНАЛИТИКА [1]

Лев Манович[2]

Изучение больших культурных данных: социальная информатика и цифровая гуманитаристика

Я определяю культурную аналитику (Cultural Analytics) как «анализ крупных коллекций и потоков культурных данных с помощью методов их компьютерной обработки и визуализации». Я разработал этот концепт в 2005 году, а в 2007-м мы основали исследовательскую лабораторию Software Studies Initiative[3] для того, чтобы начать работать над практическими проектами. Вот примеры теоретических и практических вопросов, направляющих нашу работу:

> Что значит репрезентировать «культуру» через «данные»? Какие уникальные возможности, недоступные качественным методам гуманитарных и социальных наук, предоставляет компьютерный анализ больших объемов культурных данных? Как использовать количественные методы для изучения ключевой культурной формы нашей эпохи – интерактивных медиа? Как мы можем объединить вычислительный анализ и визуализацию больших культурных данных с качественными методами, в том числе «медленным чтением»? (Иными словами, как сочетать анализ более крупномасштабных паттернов с анализом индивидуальных артефактов и их деталей?) Как при компьютерном анализе в должной мере учесть изменчивость и разнообразие культурных артефактов и процессов, не ограничиваясь «типичным» и «наиболее популярным»?

Восемь лет спустя, работа нашей лаборатории составляет лишь малую часть обширной области исследований. Тысячи исследователей уже опубликовали десятки тысяч работ с анализом паттернов в больших коллекциях культурных данных.

---

1 Перевод статьи Льва Мановича осуществлен Андреем Возьяновым по изданию: L. Manovich, The science of culture? Social computing, digital humanities and cultural analytics // *The Datafied society: social research in the age of big data*. Amsterdam University Press, Amsterdam. Режим доступа: http:// manovich. net/ index. php/ projects/ cultural-analytics-social-computing. – 2016.

2 Лев Манович (Lev *Manovich*) – теоретик новых медиа и цифровой гуманитаристики, автор The Language of New Media (2001) и Software Takes Command (2013). Профессор Высшей школы Городского университета Нью-Йорка, основатель и руководитель Лаборатории культурной аналитики.

3 Режим доступа: softwarestudies.com.

Прежде всего, это – данные, характеризующие активность в самых популярных социальных сетях (Flickr, Instagram, YouTube, Twitter и т.д.), контент, созданный пользователями (твиты, изображения, видео и т.д.) и предоставленный ими в общий доступ через эти сети, а также взаимодействие пользователей с этим контентом (линки, лайки, репосты, комментарии). Во-вторых, исследователи также начали анализировать отдельные сферы и исторические периоды в развитии культурных индустрий, такие как веб-дизайн, модную фотографию, поп-музыку XX и литературу XIX века и т.д. Эта работа ведется в двух – с недавних пор разрабатываемых – областях: социальной информатике (Social Computing) и цифровой гуманитаристике (Digital Humanities).

Какое место здесь отведено культурной аналитике? Думаю, она сохраняет свою релевантность в качестве интеллектуальной программы. Как мы увидим, цифровая гуманитаристика и социальная информатика изучают только определенные типы культурных данных, тогда как культурная аналитика этих ограничений не имеет. Мы также не собираемся выбирать между гуманитарными и социальными целями / методологией или же вносить в их отношения иерархию. Напротив, мы заинтересованы в том, чтобы совместить оба подхода к исследованиям культуры – от гуманитаристики взять фокусировку на частностях, интерпретациях и прошлом, а от социальных наук – сосредоточение на генерализованных, формальных моделях и прогнозах. В данной статье я буду обсуждать эти и другие характерные особенности развития и актуального состояния двух подходов к изучению крупных коллекций культурных данных, указывая на пока мало изведанные возможности и идеи.

Цифровые гуманитарии используют компьютеры, главным образом, для анализа исторических артефактов, созданных профессионалами. Пример тому – романы, написанные профессиональными писателями в XIX веке. Временные рамки их исследований не выходят за границы, установленные авторским правом отдельных стран. Например, в соответствии с законом об авторском праве США, произведения, опубликованные в течение последних 95 лет, автоматически защищены копирайтом. Так, например, в 2015 году все, созданное после 1920 года, является собственностью автора, если только это не цифровой контент, опубликованный с использованием лицензии Creative Commons. Я признаю необходимость уважать законы об авторском праве, но в таком случае цифровые гуманитарии сами отрезают себе путь к изучению настоящего.

Поле социальной информатики в тысячи раз больше. Здесь исследователи с учеными степенями по информатике изучают онлайн-контент, созданный пользователями, и взаимодействие с ним. Заметим, что эти исследования проводятся не только учёными-айтишниками, профессионально идентифицирующими себя с полем «Социальная информатика»[4], но также исследователями,

[4] Диапазон тем отражают программы профильных конференций, представленные, например, здесь. Режим доступа: http://cscw.acm.

работающими в других областях компьютерного знания – таких, как, например, компьютерные мультимедиа (Computer Multimedia), компьютерное зрение (Computer Vision), автоматическое распознание музыкальной информации (Music Information Retrival), автоматический анализ речи (Natural Language Processing) и сетевая наука (Web Science). Таким образом, термин «Социальная информатика» может использоваться и в качестве зонтичного термина для всех компьютерных исследований, где изучается контент и деятельность в соцсетях. Эти исследователи имеют дело с данными, полученными после 2004 года, когда социальные сети и сервисы с функцией обмена медиа контентом начали приобретать популярность. (Поскольку проведение исследования и публикация статьи занимают один-два года, то в публикациях 2015 года, как правило, используют данные, собранные в 2012-2014 гг.) Коллекции данных здесь обычно гораздо больше, чем в цифровой гуманитаристике. Десятки или даже сотни миллионов постов, фотографий или других элементов – не редкость. Поскольку большинство пользовательского контента создается обычными людьми, а не профессионалами, социальная информатика по умолчанию изучает непрофессиональную, вернакулярную культуру.

Масштаб этого исследовательского поля способен удивить гуманитариев и людей искусства, которые могут и не подозревать, сколько народа трудится в информатике и в смежных с нею областях. Например, поиск в Google Scholar по запросу «алгоритм составления коллекций данных в Twitter» (“twitter dataset algorithm”) выдал 102 000 статей, запрос «коллекции видео-данных YouTube» (“YouTube video dataset”) выдал 27 800 статей, а поиск «алгоритм обработки изображений Flickr» (“flickr images algorithm”) – 17 400[5]. Поисковой запрос «коллекции данных, составленные на основе алгоритмов автоматического распознания эстетически привлекательных изображений» (“computational aesthetics dataset”) дал 14 100 результатов[6]. Даже если действительные цифры гораздо меньше, они все равно впечатляют. Очевидно, не все публикации затрагивают вопросы, прямо адресованные культуре, но многие делают это.

В таблице, размещенной ниже, приведены различия между этими двумя областями научного знания, как я их вижу:

org/2016/submit/ (автор дает ссылку на сайт конференции CSCW-2016, Computer-Supported Cooperative Work and Social Computing – *прим. переводчика*).

5 Режим доступа: https://scholar.google.com.

6 Режим доступа: https://scholar.google.com.

Области исследований	Социальная информатика и различные области информатики, где изучаются социальные сети и цифровые среды для обмена контентом	Цифровая гуманитаристика (в частности, исследования в ЦГ, в рамках которых производится количественный анализ с использованием методов информатики)
Количество публикаций	Десятки тысяч	Менее 100
Исследуемые период и материал	Веб-сайты, контент и активность пользователей в социальных медиа, начиная с 2004 г.	Исторические артефакты до начала XX века
Создатели изучаемых артефактов	Обычные люди, которые предоставляют контент для общего доступа в социальных сетях	Профессиональные писатели, художники, композиторы и т.д.
Размер коллекций данных (dataset)	От тысяч до сотен миллионов объектов и миллиарды отношений между ними	Обычно сотни или тысячи элементов

Почему исследователи-компьютерщики редко работают с большими коллекциями исторических данных? Как правило, они обосновывают необходимость своих исследований, ссылаясь на уже внедренные в промышленных масштабах разработки. Например, используют системы поиска по онлайн-контенту или информационные фильтры для него. Предполагается, что информатика создаст улучшенные алгоритмы и другие вычислительные технологии, полезные для промышленности и государственных организаций. Анализ исторических артефактов выпадает из этого круга задач, и, следовательно, мало кто из исследователей-айтишников работает с историческими данными (за исключением одной единственной области – цифрового наследия (Digital Heritage)).

Однако при взгляде на многие работы по информатике становится ясно, что в действительности здесь занимаются гуманитаристикой или исследованиями коммуникации (в привязке к современным медиа), только в большем масштабе. Возьмём, например, эти недавние публикации: «Подсчитывая визуальные предпочтения жителей планеты» и «Что мы выкладываем в Instagram: первый анализ фотоконтента и типов пользователей Instagram»[7].

[7] См. K. Reinecke, K. Gajos: Quantifying Visual Preferences Around the World, in: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, N. Y.: ACM, 2014, 11-20, Режим доступа: http://www.eecs.harvard.edu/~kgajos/papers/2014/reinecke14visual.pdf; Y. Hu, L. Manikonda, S. Kambhampati: What We Instagram: A First Analysis of Instagram Photo Content and User Types, in: *Proceedings of the 8th*

В первом исследовании на основе 2,4 млн оценок, полученных от 40 000 человек из 179 стран, анализируются предпочтения в вебдизайне сайтов по всему миру. Очевидно, изучение эстетики и дизайна традиционно было областью гуманитарных наук. Во втором исследовании проанализированы наиболее частотные темы на фотографиях из Instagram – проблематика, сопоставимая с изучением жанров в голландской живописи XVII века в рамках истории искусств.

Другой пример – статья, озаглавленная «Что такое Twitter – социальная сеть или новостное СМИ?».[8] Опубликованная в 2010 году, с тех пор она была процитирована 3284 раза в других публикациях по информатике [9]. Это был первый масштабный анализ социальной сети Twitter на базе 106 млн твитов от 41,7 млн пользователей, исследование, где, в частности, рассматривались лидирующие по популярности темы и было показано, «на какие категории можно разделить лидирующие темы, как долго они остаются на вершине популярности, и какое количество пользователей участвуют в формировании этого рейтинга». Это – классические для исследований коммуникации вопросы, возвращающие к пионерской работе Пола Ф. Лазарсфельда и его коллег, которые в 1940-е вручную пересчитывали темы радиопередач. Но учитывая, что Twitter и другие микроблоговые сервисы представляют собой новую форму медиа – как масляная живопись, печатные книги или фотографии в прошлом, – то понимание медийной специфики Twitter тоже является темой для гуманитарных наук.

Незначительное число публикаций находятся на пересечении цифровой гуманитаристики и социальной информатики. В них используются вычислительные методы и алгоритмы, разработанные цифровыми исследователями для изучения современного пользовательского контента и медиа, но применяются эти методы и алгоритмы к историческим артефактам, созданным профессионалами (т.е. профессиональными художниками, писателями, издателями, музыкантами или режиссерами). Выдающиеся примеры – такие тексты, как «На пути к автоматизированному выявлению влияний в искусстве»[10], «Вирусные тексты: моделирование повторного использования текста в газетах XIX века»[11], «Измеряя количе-

---

International AAAI Conference on Weblogs and Social Media. ICWSM, 2014. Режим доступа: http://rakaposhi.eas.asu.edu/instagram-icwsm.pdf.

8 H. Kwak, Ch. Lee, H. Park, S. Moon: What is Twitter, a Social Network or a News Media? in: *Proceedings of the 19th International World Wide Web (WWW) Conference.* ACM, 2014, 591-600. Режим доступа: http://www.eecs.wsu.edu/~assefaw/CptS580-06/papers/2010-www-twitter.pdf.

9 Режим доступа: https://scholar.google.com/citations?user=M6i3Be0AAAAJ&hl=en.

10 B. Saleh, K. Abe, R. Singh, A. A. Elgammal: Toward Automated Discovery of Artistic Influence," *Multimedia Tools and Applications* (Springler, 8/19/2014): 1-27. Режим доступа: http://arxiv.org/abs/1408.3218.

11 D. Smith, R. Cordell, E. Dillon: Infectious texts: Modeling text reuse in nineteenth-century newspapers, in: *Proceedings of 2013 IEEE Conference*

ственные характеристики эволюции современной западной поп-музыки»[12] и «Быстрее, динамичнее, темнее: изменения в голливудском кино за 75 лет»[13].

Еще несколько лет назад единственный проект, где культурная история изучалась по-настоящему крупномасштабно, на миллионах текстов, был реализован, скорее, представителями точных наук, нежели гуманитариями. Я имею в виду N-Gram Viewer, который был создан в 2010 году научными сотрудниками Google Джоном Оруэнтом и Уиллом Брокманом на основе прототипа, разработанного биологом и прикладным математиком, гарвардскими аспирантами. Тем не менее, в последнее время мы наблюдаем увеличение объемов данных, с которыми работают цифровые гуманитарии. Например, в проекте «Картирование изменяющихся жанров в антологиях»[14] литературовед Тед Андервуд и его соавторы проанализировали 469 200 томов из цифровой библиотеки Trust Digital Library[15]. Искусствовед Максимилиан Ших и его коллеги изучили жизненные траектории 120 000 выдающихся исторических деятелей (проект «Сетевая основа культурной истории»)[16]. В сфере исследований литературы, фотографии, кино и телевидения становятся доступными еще более крупные исторические коллекции данных, но эти массивы еще только предстоит проанализировать. В 2012 году Нью-Йоркский Муниципальный архив опубликовал 870 000 оцифрованных исторических фотографий Нью-Йорка[17]. В 2015-м HathiTrust предоставил исследователям доступ к данным из 4 801 237 томов, содержащих 1,8 млрд. страниц[18]. В том же году The Associated Press[19] и British Movietone[20] загрузили на YouTube 550 000

---

on Big Data. IEEE, 2013, 84-94. Режим доступа: http://www.ccs.neu.edu/home/dasmith/infect-bighum-2013.pdf.

12 J. Serrà, Á. Corral, M. Boguñá, M. Haro, J. Arcos: Measuring the Evolution of Contemporary Western Popular Music, in: Nature Scientific Reports 2, article 521 (2012). Режим доступа: http://www.nature.com/articles/srep00521.

13 J. Cutting, K. Brunick, J. DeLong, C. Iricinschi, A. Candan: Quicker, faster, darker: Changes in Hollywood film over 75 years, in: *i-Perception,* 2 (2011), 569 – 576. Режим доступа: http://people.psych.cornell.edu/~jec7/pubs/iperception.pdf.

14 Режим доступа: http://arxiv.org/abs/1309.3323.

15 T. Underwood, M. Black, L. Auvil, B. Capitanu: Mapping Mutable Genres in Structurally Complex Volumes, in: *Proceedings of the 2013 IEEE Conference on Big Data.* IEEE, 2013. Режим доступа: http://arxiv.org/abs/1309.3323.

16 M. Schich, Ch. Song, Y.-Y. Ahn, A. Mirsky, M.Martino, A.-L. Barabási, D. Helbing: A network framework of cultural history, in: Science, 345 (2014), 558-562. Режим доступа: http://www.uvm.edu/~cdanfort/csc-reading-group/schich-science-2014.pdf.

17 Режим доступа: http://www.theatlantic.com/photo/2012/04/historic-photos-from-the-nyc-municipal-archives/100286/.

18 Режим доступа: https://sharc.hathitrust.org/features, retreived 8/20/2015.

19 Режим доступа: https://youtube.com/c/apachive.

20 Режим доступа: https://www.youtube.com/channel/UCHq777_waKMJw6SZdABmyaA

оцифрованных новостных сообщений за период с 1895 года по сегодняшний день[21].

Почему важно иметь в своем распоряжении такие крупные коллекции культурных данных? Почему бы просто не обойтись меньшими выборками? Я считаю, на то есть причины. Во-первых, если мы хотим получить репрезентативную выборку, то должны сначала располагать гораздо большим набором единиц, из которого можно выбирать, или, по крайней мере, хорошим пониманием того, *что* включает в себя такой набор. Так, например, если мы хотим создать репрезентативную подборку фильмов XX века, мы можем использовать IMDb, которая содержит информацию о 3,4 млн фильмов и телевизионных шоу, включая отдельные эпизоды[22]. Аналогичным образом, мы можем создать хорошую подборку полос старых американских газет, используя Историческую коллекцию американских газет, включающую миллионы оцифрованных газетных полос из Библиотеки Конгресса[23]. Но во многих других областях культуры таких больших коллекций данных не существует, а без них невозможно создать репрезентативные выборки.

Вот и вторая причина. Предположив, что мы можем собрать репрезентативную выборку того или иного культурного поля, мы можем использовать её для обнаружения общих для этого поля тенденций и паттернов. Так, в уже упомянутой работе «Что мы выкладываем в Instagram: первый анализ фотоконтента и типов пользователей Instagram»[24] трое исследователей изучили тысячу фотографий в Instagram и выделили восемь наиболее частотных категорий (селфи, друзья, мода, еда, гаджеты, деятельность, домашние животные, фотографии с подписями). Выборка в тысячу фотографий была случайным образом составлена из большего набора фотографий, размещенных в публичном доступе 95 343 уникальными пользователями. Вполне возможно, эти восемь категорий были самыми популярными среди Instagram-фотографий, размещенных из разных точек планеты в то время, когда проводилось исследование. Однако, как мы увидели в наших проектах, посвященных анализу Instagram-снимков из разных городов и их частей (например, из центра Киева во время украинской революции 2014 года в проекте «Исключительное и повседневное»[25]), люди размещают фотографии других видов. В зависимости от времени и места некоторые виды фотографий могут обогнать по популярности восьмерку лидеров. Другими словами, хотя маленькая выборка и позволяет найти «типичное» или «самое популярное», она не раскрывает того, что я называю «островами контента» – разновидно-

[21] Режим доступа: http://www.ap.org/content/press-release/2015/ap-makes-one-million-minutes-of-history-available-on-youtube.

[22] IMDb. Режим доступа: http://www.imdb.com/stats.

[23] Режим доступа: http://chroniclingamerica.loc.gov/about/

[24] Reinecke, op.cit.

[25] L. Manovich, M. Yazdani, A. Tifentale, J. Chow: The Exceptional and the Everyday: 144 hours in Kyiv, 2014. Режим доступа: http://www.the-everyday.net/.

стей связного контента со специфическими семантическими и/или эстетическими характеристиками, размещаемого в публичном доступе в небольших количествах.

Можем ли мы изучить всё?

Когда я впервые начал думать о культурной аналитике в 2005 году, оба исследовательских поля – цифровая гуманитаристика и социальная информатика – только-только начинали формироваться. Я почувствовал потребность в этом новом термине, чтобы сигнализировать – работа нашей лаборатории не будет просто частью цифровой гуманитаристики или социальной информатики, но будет охватывать предметные области, изучаемые обеими. Подобно цифровым гуманитариям, мы заинтересованы в анализе исторических артефактов, но нам в равной степени интересна современная цифровая визуальная культура (в частности, Instagram). Кроме того, нам интересны как профессиональная культура, так и артефакты, созданные увлеченными непрофессионалами, художниками, находящимися за пределами мира искусства (например, deviantart.com, «крупнейшая социальная интернет-сеть для художников и почитателей искусства»[26]), или теми, кто занимается творчеством от случая к случаю (например, теми, кто иногда загружают свои фото в соцсети). Как и цифровых социологов вместе с учеными-айтишниками, нас привлекает изучение общества через социальные медиа и изучение социальных явлений, характерных для соцсетей.

Пример изучения общества через социальные медиа – поиск в городе схожих соседских сообществ (neighborhoods) на основе изучения активности в социальных медиа, как в проекте «Цифровые соседства (Livehoods): использование социальных медиа для понимания городской динамики»[27]. Пример изучения социального измерения соцсетей – анализ паттернов распространения информации в Интернете, как в проекте «Отложенные информационные каскады во Flickr: измерение, анализ и моделирование»[28]. Однако, если социальная информатика фокусируется на *социальном* в социальных сетях, то культурная аналитика сосредотачивается на *культурном*. Таким образом, наиболее релевантной областью социальных наук для культурной аналитики является социология культуры, и только потом – социология с экономикой.

---

26 Режим доступа: http://about.deviantart.com/.

27 J. Cranshaw, R. Schwartz, J. Hong, N. Sadeh: The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City, in: *The 6th International AAAI Conference on Weblogs and Social Media.* Dublin, 2012. Режим доступа: https://s3.amazonaws.com/livehoods/livehoods_icwsm12.pdf.

28 M. Cha, F. Benevenuto, Y.-Y. Ahn, K. Gummadi: Delayed information cascades in Flickr: Measurement, analysis, and modeling, in: *Computer Networks* 56 (2012), 1066–1076. Режим доступа: http://200.131.208.43/bitstream/123456789/2022/1/ARTIGO_DelayedInformationCascades.pdf.

Мы верим, что контент интернет-сайтов и соцсетей вместе с активностями пользователей дают беспрецедентную возможность для описания, построения теоретических моделей и компьютерного моделирования глобального культурного универсума, покуда продолжаются переосмысление и ревизии основополагающих концептов и инструментов гуманитаристики, разработанных для анализа «малых культурных данных» (т.е. очень избирательно сформированных и нерепрезентативных выборок). Согласно весьма авторитетному определению, данному британским критиком Мэтью Арнольдом (1869), культура – это «лучшее, что было помыслено и сказано в этом мире»[29]. Академическая гуманитаристика по большей части следовала этому определению. И когда гуманитарии восстали против собственных же канонов, включив в обновленное понимание культуры работы исключенных ранее людей (женщин, не-белых, незападных авторов, квир-людей и т.д.), то зачастую они включали лишь «лучшее» из созданного теми, кто был ранее исключен.

Культурная аналитика интересуется *всем, создаваемым всеми*. В этом смысле мы подходим к культуре так же, как лингвисты – к изучению языков, а биологи – к изучению жизни на Земле. В идеале мы хотим изучать каждое проявление культуры, а не избирательно сформированные выборки. (Эта более систематическая перспектива не так уж отличается от той, что задана культурной антропологией). Пример более широкой рамки, позволяющей сочетать профессиональное и общеупотребительное, историческое и современное, дает серия проектов, над которыми наша лаборатория работает с 2008 года. Мы изучили исторический, созданный профессионалами, культурный контент со всех обложек журнала *Time* (1923-2009); живопись Винсента Ван Гога, Пита Мондриана и Марка Ротко; 20 000 фотографий из коллекции Музея современного искусства в Нью-Йорке (MoMA); один миллион страниц из 883 комиксов манга, опубликованных за последние 30 лет. Наши исследования актуального вернакулярного контента включают *Phototrails* (сравнение визуальных профилей 13 мегаполисов на основе 2,3 млн фото в Instagram)[30]; *Исключительное и обыденное: 144 часа в Киеве* (анализ Instagram-изображений, выложенных в Киеве во время украинской революции 2014 года)[31] и *На Бродвее* (интерактивная инсталляция, позволяющая изучать Бродвей в Нью-Йорке на основе 40 млн. точек данных и фотографий, созданных пользователями)[32]. Мы также рассматривали современный любительский и полупрофессиональный художественный контент (один

[29] M. Arnold. *Culture and Anarchy.* London, 1869. Режим доступа: http://www.library.utoronto.ca/utel/nonfiction_u/arnoldm_ca/ca_all.html.

[30] N. Hochman, L. Manovich, J. Chow: Phototrails, 2013. Режим доступа: http://phototrails.net/.

[31] L. Manovich, M. Yazdani, A. Tifentale, J. Chow: *The Exceptional and the Everyday: 144 hours in Kyiv,* 2014. Режим доступа: http://www.the-everyday.net/.

[32] D. Goddemeyer, M. Stefaner, D. Baur, L. Manovich: *On Broadway,* 2014. Режим доступа: http://on-broadway.net/.

миллион работ, размещенных тридцатью тысячами полупрофессиональных художников на www.deviantart.com). В настоящее время мы изучаем набор данных из 265 млн. изображений, выложенных в Twitter по всему миру в 2011–2014 гг. Подытоживая, мы не проводим границу между (небольшими) историческими артефактами, созданными профессионалами, и (более объемным) цифровым контентом, созданным непрофессионалами и размещенным в Сети. Напротив, мы свободно черпаем данные из обоих источников.

Очевидно, соцсети – это еще не все человечество, а размещаемый в них контент порой специфичен для этих сетей (как селфи для Instagram) и отличается от всего, что существовало до их появления. Форму контенту также придают технологические инструменты и интерфейсы, используемые для его создания, сбора, редактирования и совместного использования (например, фильтры в Instagram или шаблоны коллажей, предлагаемые сторонними приложениями – такими как InstaCollage). Доступные пользователю культурные действия также определены технологиями. Например, в соцсетях вы можете «лайкнуть», поделиться и/или прокомментировать контент. Иными словами, как и в квантовой физике, инструмент здесь может влиять на явления, которые мы хотим изучить. Все это необходимо самым тщательным образом учитывать, когда мы изучаем контент, созданный пользователями, и их активности в соцсетях. Хотя при помощи API легко получить доступ к большим объемам контента в соцсетях, это не «всё» от «всех». (API означает Application user interface, пользовательский интерфейс вебприложения. Это механизм, который позволяет любому загрузить большие объемы пользовательского контента из всех основных социальных сетей. Во всех публикациях по информатике для загрузки данных, подлежащих анализу, используется API.)

Общее и особенное

Когда гуманитарные науки занимались «малыми данными» (контент, созданный отдельными авторами или небольшими группами), социологическая перспектива была лишь одной из многих возможностей интерпретации – если только вы не были марксистом. Но как только мы начинаем изучать онлайн-контент и сетевое поведение миллионов людей, эта перспектива становится почти неизбежной. В случае больших культурных данных *культурное* и *социальное* накладываются друг на друга. Большие группы людей из разных стран и социально-экономических слоев (социологическая перспектива) делятся фотографиями, видео, текстами, тем самым совершая эстетический выбор (перспектива гуманитарных наук). Из-за этого наложения вопросы, изучаемые в социологии культуры XX века (пример тому – ее самый влиятельный представитель Пьер Бурдье[33]), имеют прямое отношение к культурной аналитике.

[33] P. Bourdieu: *Distinctions. A Social Critique of the Judgment of Taste.* Harvard University Press, 1984.

Принимая во внимание, что некоторые демографические категории стали само собой разумеющимися для нашего способа размышлять об обществе, сегодня вполне естественно группировать людей по этим категориям и сравнивать их по социальным, экономическим или культурным показателям. Например, Pew Research Center регулярно размещает статистические отчеты об использовании популярных социальных платформ, разбивая выборку пользователей по таким демографическим признакам, как пол, этническая принадлежность, возраст, образование, доход и место проживания (город, пригород и сельская местность)[34]. Таким образом, если нас интересуют различные детали поведения пользователей в социальных медиа – такие, как размещаемые и предпочитаемые типы снимков, используемые фильтры или позы для селфи, это логично – изучать различия между выходцами из разных стран, социально-экономических слоев, пользователей разной этнической принадлежности или технической грамотности. Ранние исследования в области социальной информатики не учитывали такие различия (в большинстве работ в области социальной информатики их не рассматривают до сих пор), принимая всех пользователей за один недифференцированный массив «человечества». Но в последнее время начинают попадаться публикации, где пользователей разбивают на демографические группы. Хотя это очень позитивная тенденция, возможно, нам не стоит здесь заходить далеко. Гуманитарное исследование процессов и явлений культуры с использованием количественных методов не должно сводиться к социологии, т.е. учету одних лишь характеристик и сетевого поведения социальных групп.

Социологическая традиция в большей степени связана с поиском и описанием *общих* паттернов в поведении человека, чем с анализом или прогнозированием поведения индивидов. Культурная аналитика также интересуется паттернами, которые могут быть выведены в ходе анализа больших коллекций культурных данных. Тем не менее, в идеале *анализ общих паттернов приведет нас к конкретным случаям,* т. е. отдельным создателям контента, их творениям или формам культурного поведения. Например, компьютерный анализ всех фотографий, сделанных фотографом за время долгой карьеры, может вывести нас на маргинальные случаи – фотографии, которые в наибольшей степени выпадают из общего ряда. Так же мы можем анализировать миллионы Instagram-снимков, выложенных в сеть во множестве городов, чтобы обнаружить типы изображений, присущие каждому мегаполису (пример из исследования, в настоящий момент выполняемого нашей лабораторией).

Другими словами, мы можем сочетать озабоченность социальных наук (и науки в целом) *общим* и *регулярным* с озабочен-

[34] Demographics of Key Social Networking Platforms, 2015. Режим доступа: http://www.pewinternet.org/2015/01/09/demographics-of-key-social-networking-platforms-2/

ностью гуманитаристики *индивидуальным* и *частным* (в конце концов, все великие художники в истории искусства были чужаками для современников). Только что описанные примеры анализа массивных коллекций данных с аналитическим приближением (зуммированием) отдельных объектов из этих коллекций, иллюстрирует один, но далеко не единственный из способов достичь этого сочетания.

Наука о культуре?

Цель науки – объяснить явления и представить компактные математические модели, которые описывают, как эти явления работают. Три закона Ньютона в физике – прекрасный пример того, как классическая наука достигала этой цели. С середины XIX века целый ряд новых областей научных исследований принял новый, вероятностный подход. Первым примером было статистическое распределение, описывающее наиболее вероятные скорости частиц газа, – его представил Максвелл в 1860 году (в настоящее время оно называется распределением Максвелла-Больцмана). А как насчет социальных наук? На протяжении XVIII и XIX веков многие мыслители ожидали, что со временем будут найдены законы, управляющие обществом, аналогичные физическим[35]. Этого так и не произошло (ближе всего социально-философская мысль XIX века подошла к формулированию объективных законов в работах Карла Маркса). Вместо этого позитивистские науки об обществе, возникшие в конце XIX – начале XX века, освоили вероятностный подход. Таким образом, вместо того чтобы искать законы, детерминирующие жизнь общества, социологи изучают корреляции между измеримыми параметрами и моделируют отношения между «зависимыми» и «независимыми» переменными при помощи статистических методов. За детерминистской и вероятностной парадигмами в науке последовала парадигма математического моделирования – создание рабочих компьютерных моделей для имитации поведения систем. Первая масштабная компьютерная симуляция была создана в 1940 году в Манхэттенском проекте для моделирования ядерного взрыва. Впоследствии симуляция была адаптирована во многих точных науках, а в 1990-е годы воспринята социологами.

В начале XXI века объемы цифрового контента и взаимодействия пользователей в Сети позволяют нам помыслить «точную науку о культуре». Так, летом 2015 года пользователи Facebook размещали 400 миллионов фотографий и отправляли 45 миллиардов сообщений в день[36]. Этот порядок величин по-прежнему намного меньше, чем у атомов и молекул – например, 1 $см^3$ воды содержит $3{,}33*10^{22}$ молекул. Тем не менее, это больше, чем число нейронов

[35] Ph. Ball: *Critical Mas*s. L: Arrow Books, 2004, 69-71.

[36] Режим доступа: http://expandedramblings.com/index.php/by-the-numbers-17-amazing-facebook-stats/15/.

в нервной системе среднестатистического взрослого человека, оцениваемое в 86 миллиардов. Но если наука сегодня располагает несколькими фундаментальными подходами к изучению и пониманию явлений – это детерминистские законы, статистические модели и компьютерное моделирование, – какие из них гипотетическая наука о культуре должна адаптировать для своих нужд?

Из работ исследователей-айтишников, изучающих социальные медиа на основе коллекций данных, ясно, что по умолчанию в них используется статистический подход[37]. Содержание социальных медиа и поведение пользователей они описывают в терминах вероятностей. Этот подход предполагает создание статистических моделей – то есть математических уравнений, определяющих отношения между переменными, которые могут быть описаны, скорее, с помощью вероятностных распределений, чем конкретных значений. В большинстве работ сегодня также используется контролируемое машинное обучение – автоматическое создание моделей, способных классифицировать или предсказывать значения новых данных на основании уже имеющихся. В обоих случаях модель учитывает лишь часть данных, что типично для статистического подхода.

Исследователи-айтишники, изучающие социальные медиа, используют статистику иначе, чем социологи. Социологи хотят *объяснить* социальное, экономическое или политическое явление (например, влияние социального положения семьи на детскую успеваемость). Айтишники, как правило, не обременяют себя объяснением паттернов, обнаруженных в социальных медиа, через установление их связи с некими внешними социальными, экономическими или технологическими факторами. Вместо этого они обычно либо изучают явления, имманентно присущие социальным медиа, либо пытаются предсказать внешние явления, используя информацию, извлеченную из социальных медиа. Пример первого подхода – статистическое описание того, сколько в среднем добавлений в «Избранное» может получить фото на Flickr по истечении определенного периода времени[38]. Пример второго подхода – сервис Google Flu Trends, который предсказывает активность вируса гриппа, комбинируя статистику поиска в Google с официальными эпидемическими данными от CDC (Центра США по контролю и профилактике заболеваний)[39].

Разница между детерминистскими законами и недетерминистскими моделями заключается в том, что последние описывают ве-

---

37 Ученые-айтишники также используют много новых методов, которые не были частью статистики прошлого века, включая глубинный анализ данных (data mining) и машинное обучение. Я обсуждаю эти различия в статье Data Science and Digital Art History, in: *International Journal for Digital Art History*, 1 (2015), Режим доступа: https://journals.ub.uni-heidelberg.de/index.php/dah/article/view/21631.

38 Cha, op. cit.

39 Режим доступа: http://googleresearch.blogspot.com/2014/10/google-flu-trends-gets-brand-new-engine.html, 10/31/2014.

роятности, но не определенности. Законы классической механики применимы к любым макроскопическим объектам. И напротив, вероятностная модель для прогнозирования числа отметок «Избранное» на Flickr, рассматриваемая в качестве производной от времени, прошедшего с момента загрузки, не может дать точную цифру для каждого отдельного фото. Она описывает только общую тенденцию. Кажется, это подходящий метод для «науки о культуре». Если мы вместо этого начнем постулировать детерминистские законы культурной активности человека, то что же произойдет с идеей свободы воли? Даже в случае, казалось бы, довольно автоматичного культурного поведения (добавление в «Избранное» в социальных сетях фотографий с определенным содержанием, таким как красивые пейзажи, милые питомцы или позирующие девушки) мы не хотели бы редуцировать людей к механическим автоматам для передачи мемов.

В настоящее время акцент на вероятностных моделях в изучении онлайн-поведения не оставляет места для третьей научной парадигмы – компьютерного моделирования (симуляции). Насколько мне известно, ни в социальной информатике, ни в цифровой гуманитаристике компьютерное моделирование еще не освоено в качестве инструмента для изучения пользовательского контента, его тематики, типов изображений и т.д. Если в 2009 году ученые из исследовательского центра Almaden IBM создали компьютерную модель визуальной коры головного мозга человека с помощью 1,6 миллиарда виртуальных нейронов с 9 триллионами синапсов[40], то почему мы не можем помыслить модель, например, всего контента, ежегодно производимого пользователями Instagram? Или всего контента, размещаемого пользователями крупнейших социальных сетей? Или типов изображений, размещаемых людьми для общего доступа? Смысл таких симуляций будет не в том, чтобы построить правильную модель или точно предсказать, чем пользователи будут делиться в Сети в следующем году. Вместо этого мы можем последовать совету авторов влиятельного учебника «Компьютерное моделирование для социальных исследователей», которые утверждают, что одна из целей моделирования – «добиться лучшего *понимания* некоторых особенностей социального мира» и что моделирование может быть методом «*построения теории*»[41] (выделено мной. – Л. М.). Так как компьютерное моделирование требует разработки эксплицированной и точной модели явления, осмысление возможностей моделирования культурных процессов может помочь нам в разработке более четких и подробных теорий, чем те, что мы обычно используем (для примера того, как агентное моделирование может быть использовано для изучения эволюции че-

[40] Режим доступа: http://www.popularmechanics.com/technology/a4948/4337190/, 12/17/2009.

[41] N. Gilbert, K. Troitzsch: *Simulation for the Social Scientist*, McGraw Hill Education, 2005, 3-4.

ловеческих обществ, см. проект «Война, пространство и эволюция сложных обществ Старого Света»[42]).

А как насчет «больших данных»? Разве они не представляют собой новую парадигму со своими собственными методами исследования? Это сложный вопрос, заслуживающий отдельной статьи. (Если мы говорим о методах и техниках исследования, то развитие компьютерной сферы в 2000-х годах, в частности, увеличение мощности процессора и объема оперативной памяти, а также использование графических процессоров и вычислительных кластеров, было, вероятно, важнее, чем доступность больших объемов данных. И даже если в использовании машинного обучения с большими тренировочными массивами данных достигнуты значительные успехи, в большинстве случаев это само по себе не обеспечивает объяснения явлений.) Тем не менее, в качестве заключения я хочу упомянуть одно интересное для гуманитариев понятие, которое мы можем заимствовать из аналитики больших данных и использовать его несколько в ином направлении.

Социальная наука XX века работала над тем, что мы можем назвать «длинными данными»[43]. То есть число случаев, как правило, во много раз превосходило число анализируемых переменных. Представьте, что мы расспросили 2000 человек об их доходах, уровне образования в семье и периоде обучения. В результате мы имеем 2000 случаев и три переменные. Мы можем изучить корреляции между этими переменными, кластеризовать данные или выполнять другие виды статистического анализа.

Самый экстремальный перекос в эту сторону наблюдался у истоков социальной науки. Первый позитивистский социолог – Карл Маркс – делил все человечество на два класса: людей, владеющих средствами производства, и людей, не владеющих ими, т.е. на капиталистов и пролетариат. Позже социологи добавили другие разделения. Сегодня эти классификации присутствуют в многочисленных опросах, исследованиях, отчетах, в популярных СМИ и в научных публикациях – чаще всего это пол, раса, этническая принадлежность, возраст, образование, доход, место проживания, вероисповедание и некоторые другие (список дополнительных переменных варьируется от исследования к исследованию). Но независимо от деталей данные, собранные, проанализированные и интерпретированные, – очень «длинные». Целые популяции или выборки, составленные на их основе, описываются с помощью малого числа переменных.

---

42 P. Turchina, T. Currieb, E. Turnerc, S. Gavriletsd: War, space, and the evolution of Old World complex societies, in: *PNAS,* 110/41 (2013), 16384-16389.

43 Я использую этот термин иным образом, нежели Сэмьюэл Абресман в S. Abresman: Stop Hyping Big Data and Start Paying Attention to 'Long Data'// *Wired*, 1/29/2013. Режим доступа: http://www.wired.com/2013/01/forget-big-data-think-long-data/.

Но почему это так? В области компьютерного анализа медиа и компьютерного зрения исследователи-компьютерщики используют алгоритмы для извлечения тысяч характеристик каждого изображения, видео, твита, имейла, и т.д.[44]. И хотя, например, Винсент Ван Гог создал всего 900 картин, они могут быть описаны по тысячам независимых параметров. Точно так же мы можем описать всех и каждого жителя большого города по миллионам независимых показателей, извлекая всевозможные характеристики из поведения в социальных медиа. В качестве другого примера возьмем наш собственный проект «На Бродвее», где мы представляем Бродвей на Манхэттене через 40 миллионов точек данных и снимков, используя сообщения, изображения и регистрации в сети (check-ins), привязанные к этой улице в Twitter, Instagram и Foursquare, а также данные о поездках на такси и показатели из переписи населения США по прилегающим районам[45]. Другими словами, вместо *длинных данных* мы можем иметь в своем распоряжении *широкие данные* – очень большое и потенциально бесконечное число переменных для описания набора случаев. Обратите внимание, что, если у нас переменных больше, чем случаев, такое представление будет противоречить здравому смыслу и социальных наук, и науки о данных. Последняя отсылает к процессу усиления управляемости большим количеством переменных через *снижение размерности*. Но «широкие данные» дают нам возможность переосмыслить фундаментальные допущения насчет того, что такое общество и как его изучать; что такое культура, карьера в искусстве, совокупность изображений, группа людей с похожими эстетическими вкусами и так далее. Вместо того, чтобы выделять культурную историю с использованием одного параметра (время), двух (время и географическое расположение) или еще нескольких (например, тип медиа и жанр), мы можем задействовать бесконечное число параметров. Целью «анализа широких данных» будет не только поиск новых сходств, взаимосвязей и кластеров в универсуме культурных артефактов, но, в первую очередь, помощь в проблематизации нашего привычного взгляда на вещи, где определенные параметры считаются само собой разумеющимися. И это один из примеров главного метода культурной аналитики – *остранения* [46], «взгляда с удивлением» на наши основные культурные концепты, а также способы организации и понимания коллекций культурных данных; когда данные и методы их обработки используются для вопрошания о том, как мы думаем, видим и что в конечном счете делаем с нашими знаниями.

---

44 Я объясняю причину использования большого количества функций в статье Data Science and Digital Art History, op. cit.

45 Режим доступа: http://www.on-broadway.nyc/.

46 Термин «остранение» ввел русский литературный теоретик Виктор Шкловский в своем эссе «Искусство как прием» в 1917 году. Режим доступа: http://www.opojaz.ru/manifests/kakpriem.html

Благодарности

Я благодарен моим коллегам по исследованиям в области информатики и цифровой гуманитаристики за многочисленные дискуссии на протяжении долгих лет. Моя благодарность также адресована студентам, докторантам и исследователям, которые работали в нашей лаборатории с 2007 года и так многому меня научили. Наша работа была щедро поддержана Фондом Эндрю Меллона, Национальным фондом развития гуманитарных наук, Национальным научным фондом, Национальным научно-исследовательским вычислительным центром энергетики (NERSC), отделением аспирантуры Городского университета Нью-Йорка (CUNY), Калифорнийским институтом телекоммуникаций и информационных технологий (Calit2), Университетом Калифорнии – Сан-Диего (UCSD), Гуманитарным исследовательским институтом Калифорнии, Министерством образования Сингапура и Музеем современного искусства (Нью-Йорк).

Перевод с английского Андрея Возьянова