

Дата-грамотность и наука о данных: Образовательные подходы и решения

Deryabin, Andrey; Popov, Alexandr; Gluhov, Pavel

Preprint / Preprint

Arbeitspapier / working paper

Empfohlene Zitierung / Suggested Citation:

Deryabin, A., Popov, A., & Gluhov, P. (2020). *Дата-грамотность и наука о данных: Образовательные подходы и решения*. Moscow: Russian Presidential Academy of National Economy and Public Administration (RANEPA), Federal State Budgetary Educational Institution of Higher Education. <https://nbn-resolving.org/urn:nbn:de:0168-ssoar-83009-0>

Nutzungsbedingungen:

Dieser Text wird unter einer CC BY-NC-ND Lizenz (Namensnennung-Nicht-kommerziell-Keine Bearbeitung) zur Verfügung gestellt. Nähere Auskünfte zu den CC-Lizenzen finden Sie hier:

<https://creativecommons.org/licenses/by-nc-nd/4.0/deed.de>

Terms of use:

This document is made available under a CC BY-NC-ND Licence (Attribution-Non Commercial-NoDerivatives). For more information see:

<https://creativecommons.org/licenses/by-nc-nd/4.0>

**Федеральное государственное бюджетное образовательное
учреждение высшего профессионального образования
«РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА
И ГОСУДАРСТВЕННОЙ СЛУЖБЫ ПРИ ПРЕЗИДЕНТЕ
РОССИЙСКОЙ ФЕДЕРАЦИИ»**

Дерябин А.А., Попов А.А., Глухов П.П.

**Дата-грамматность и наука о данных:
Образовательные подходы и решения**

Москва 2020

Аннотация:

В настоящее время в литературе описано нет так много инициатив по разработке образовательных программ и преподаванию науки о данных учащимся подросткового и юношеского возраста. В контексте фундаментальных изменений в экономике и на рынке труда важное значение приобретает внедрение образовательных программ в области анализа данных и машинного обучения на всех уровнях образования с приоритетом интеграции математического, естественно-научного и социально-гуманитарного знания. Представлен обзор и анализ зарубежного опыта и основных дискуссионных тем в разработке образовательных модулей науки о данных и машинного обучения для подросткового и юношеского возраста. Рассматривается понятие «дата-грамотности» как одной из базовых компетентностей в датафицированном обществе и подходы к науке о данных как, с одной стороны, направлению научно-инженерного образования, а с другой как к социокультурной практике и инструменту гражданской активности. Развитие предметных навыков учащихся соотносится с компетентностными требованиями стандартной индустриальной модели исследования данных. Особое внимание уделяется таким практическим аспектам разработки содержания обучения как использование аутентичных данных в процессе обучения и обеспечение релевантности данных интересам учащихся. Приводятся характеристики учебных данных с точки зрения их сложности и эффекта на дата-грамотность, критическое мышление и предметные знания учеников. В заключении рассматриваются условия становления субъектности учащихся через культивирование активной позиции «дата-продюсера» и решение практико-ориентированных образовательных задач в противоположность традиционным дисциплинарным образовательным подходам. Делается вывод о необходимости преподавания начал науки о данных в подростковом и юношеском возрасте как для развития общего критического мышления, так и становления основ инженерно-технической культуры учащихся в режиме пропедевтики будущей деятельности. Материалом для анализа послужили преимущественно публикации 2014-2020 годов в научных изданиях, индексируемых в базе данных Scopus.

Ключевые слова: *наука о данных; машинное обучение; образовательные программы; дата-грамотность; анализ данных; data science; искусственный*

интеллект; общее образование; профориентация; дополнительное образование; информатика; STEM; обществознание

Введение

Наука о данных (Data Science) затрагивает все больший круг сфер обыденной жизни, но общее образование в России пока еще не отреагировало на новые вызовы, связанные с этим аспектом цифровизации. Изменения в технологиях, экономике и обществе за последние два десятилетия сформировали новую повестку как для преподавания математики и информационных технологий, так и для общественных наук и медиа-образования. Образование во всех этих областях требует переосмысления содержания и методов преподавания в связи с растущим значением науки о данных и искусственного интеллекта в контексте фундаментальных изменений в экономике и на рынке труда.

Россия занимает 48-е место по готовности к цифровой экономике со значительным отрывом от лидирующих стран, таких, как Швеция, Сингапур, Нидерланды, согласно международному Рейтингу сетевой готовности¹. Этот индекс измеряет, насколько хорошо экономики стран используют цифровые технологии для повышения конкурентоспособности и благосостояния, а также оценивает факторы, влияющие на развитие цифровой экономики.

Сформировать ответ на вызов сохранения конкурентоспособности призван ряд правительственных инициатив в научно-технологической и образовательной политике, в частности, “Национальная стратегия развития искусственного интеллекта на период до 2030 года”.² Основными направлениями повышения уровня обеспечения российского рынка технологий искусственного интеллекта квалифицированными кадрами являются: (а) разработка и внедрение образовательных модулей в рамках образовательных программ всех уровней образования [...] для получения гражданами знаний, приобретения ими компетенций и навыков в области математики, программирования, анализа данных, машинного обучения, способствующих развитию искусственного интеллекта. При этом в целях развития перспективных методов искусственного интеллекта приоритетное значение приобретает конвергентное знание,

¹ Network Readiness Index 2019 [Электронный ресурс]. URL <http://www.networkreadinessindex.org> (дата обращения 10.04.2020).

² Указ Президента Российской Федерации от 10.10.2019 г. № 490 О развитии искусственного интеллекта в Российской Федерации [Электронный ресурс] / Сайт Президента РФ. URL: <http://www.kremlin.ru/acts/bank/44731> (дата обращения 10.04.2020)

обеспечиваемое в том числе за счет интеграции математического, естественно-научного и социально-гуманитарного образования.

В соответствии с этой повесткой и с учетом динамично развивающейся сферы Data Science российские образовательные программы разных уровней требуют содержательного обновления. В то же время попытки внедрить образование в сфере науки о данных на уровне общего образования достаточно редки даже в странах верхней десятки Рейтинга сетевой готовности. Настоящая статья содержит обзор основных дискуссионных тем, касающихся преподавания учащимся подросткового и юношеского возраста науки о данных и машинного обучения:

1. Цели и ожидаемые результаты преподавания науки о данных: новая грамотность vs. предметное знание и навыки. Содержание понятия «дата-грамотность». Данные как содержание коммуникации: этические и социальные аспекты. Анализ данных как социокультурная практика и инструмент гражданской активности.

2. Практические аспекты преподавания науки о данных и машинного обучения учащимся школьного возраста: сложные аутентичные vs. простые подготовленные учебные данные; релевантность данных интересам учащихся.

Термин «наука о данных» (data science) широко используется, когда речь идет об использовании компьютерных методов для сбора, обработки, анализа и визуализации больших объемов данных. При этом далеко не всегда речь идет о так называемых «больших данных» (big data), под которым понимаются данные, чей объем, многообразие и скорость прироста превосходят возможности традиционных систем хранения и обработки информации, существовавших на рубеже 2000-х годов (1). Методы науки о данных могут применяться для работы как с «большими», так и с тривиальными объемами данных.

Основные компетенции в data science и машинном обучении на уровне профессионального образования преимущественно задаются соответствующими элементами стандартного процесса по исследованию данных (CRISP-DM)³, который включает в себя следующие фазы: *понимание целей – начальное изучение данных – подготовка данных – моделирование – оценка – внедрение* (2). На более низком уровне эта модель предусматривает наличие компетенций по поиску, получению данных, их

³ CRISP-DM (Cross-Industry Standard Process for Data Mining) — межотраслевой стандартный процесс для исследования данных, проверенная в промышленности и наиболее распространённая методология по исследованию данных.

курурованию, манипулированию ими, обработке, анализу, визуализации и представлению результатов исследования в разных формах.

Дата-грамотность и ее социокультурные измерения

Учитывая то, что данные сегодня являются повсеместными и затрагивают все аспекты существования индивида в обществе (например, в связи с сбором разными агентами персональных данных и их последующем использовании при формировании персонализированных медийных сообщений), в академическом сообществе обсуждается не только преподавание науки о данных как учебной дисциплины для будущих специалистов в этой области, но и понятие «дата-грамотности» (data literacy), применимой в более широком контексте общественной жизни и этической проблематики. Дата-грамотность как способность понимать, что такое данные и на что они способны, рассматривается как важный элемент формирования личности, способной принимать рациональные решения, осуществлять этический выбор и вырабатывать политические мнения в современном цифровом мире (3).

В этом отношении понятие дата-грамотности близко к ранее вошедшим в обиход «медиа-грамотности» или «информационной грамотности», формируя наряду с ними букет современных цифровых компетенций.

Вне зависимости от того, нацелено ли образование в области науки о данных на подготовку к профессиональной деятельности или просто является сопутствующим техническим навыком, оно имплицитно формирует, с одной стороны, *рационального индивида*, понимающего, как и с какими последствиями он использует или предоставляет данные, когда использует свой телефон, читает новости, прокладывает маршрут в навигаторе, или взаимодействует с онлайн-магазином. С другой стороны, навыки анализа данных позволяют индивиду интерпретировать статистику и непредвзято судить о общественных, технических и политических процессах и повестках, формируя, таким образом *информированного гражданина* (3 р. 29). Примером может быть сложность интерпретации неполных и противоречивых статистических данных о распространении коронавируса COVID-19, которую испытывали аудитории средств массовой информации по всему миру в марте 2020 года.⁴

⁴ Например: Koerth M., Bronner L., Mithani J. Why It's So Freaking Hard To Make A Good COVID-19 Model. Published Mar. 31, 2020. URL <https://fivethirtyeight.com/features/coronavirus-case-counts-are-meaningless/>

Дата-грамотность включает в себя способность понимать и оценивать информацию, которая может быть получена на основе данных ((4); (5); (6)).

Владеющий дата-грамотностью учащийся должен владеть количественными и аналитическими инструментами, необходимыми для решения задачи и способностью применять их в контексте, чтобы анализировать, интерпретировать и сообщать другим результаты анализа данных (7).

Дата-грамотность лежит на пересечении областей количественного мышления и науки о данных. Количественное мышление (quantitative reasoning) – это способность применять математические принципы, критическое мышление и ясную логику к решению каждодневных проблем ((8); (9); (10); (11); (12)).

Мы предлагаем провести следующее различие между наукой о данных и дата-грамотностью (Таблица 1):

Таблица 1. Различение науки о данных и дата-грамотности

Дата-грамотность	Наука о данных
Способность понимать и оценивать информацию о реальном мире, которая может быть получена из данных	Умение решать с помощью данных производственные и бизнес-задачи
Критическое мышление в датафицированном мире	Технические навыки создания дата-продуктов
Одна из базовых современных грамотностей наряду с медийной и компьютерной	Область профессионализации
Общее и дополнительное образование	Профессиональное образование
Рациональный индивид и информированный гражданин	Специалист
Гражданское общество	Индустрия

Мастерство аналитика данных основывается на понимании предметного контекста, в котором находится набор данных, знании математики, статистики и на компьютерных навыках((13) ; (14)). Последнее определение, как и многие другие, отражает тенденцию сдвигать фокус внимания в сторону технических аспектов дата-грамотности. Однако данные далеко не всегда являются объективными свидетельствами естественных феноменов, независимыми от применяемых

исследователем средств их получения и намерений определенным образом их использовать (15):⁵

1. данные всегда несут в себе следы своего технического и материального происхождения (16), (17);
2. инструменты создания и обработки данных не нейтральны, но несут в себе ту или иную имплицитную теорию с встроенным намерением редуцировать сложные феномены к полезным формам данных (16), (18);
3. данные существуют потому, что кто-то для чего-то хотел их создать; они представляют собой ценность, актив и обладают силой убеждения, следовательно, существуют не только в предметных, но также в экономических и политических границах ((19) ; (20); (21).

Fotopoulou (22) находит сходство между концепциями дата- и медиа-грамотности, в которой чтение и интерпретация медиа понимается как социокультурная практика, а отправители и реципиенты рассматриваются в связи с их принадлежностью к социальным группам и сообществам. Каждое из которых, в свою очередь, располагает собственными мотивами, символическими, материальными ресурсами и интерпретативным репертуаром ((23); (24); (25)). Хотя и не в полной мере (практики и деятельности, связанные с данными, разнообразней, чем «чтение» в теории медиа), данные, как и медиа-сообщение, можно рассматривать как содержание и канал коммуникации.

Таким образом, взаимодействие с данными является социокультурной практикой, и большие данные могут нести в себе скрытые мотивы их кураторов (26) – достаточно привести в качестве примера электоральные данные или данные здравоохранения ((27); (28); (29))⁶, что концептуально позволяет применять к ним соответствующий гуманитарный методологический аппарат с оговоркой: если такой подход продемонстрирует свою продуктивность. Например, Pangrazio и Sefton-Green (30) считают, что следует осторожно относиться к данным как к «тексту» в силу того, что конкретный набор данных может претерпевать слишком много циклов обработки (от чистки «сырых» данных к искусственной инженерии нужных переменных, и далее к метаданным) и использования его разными акторами в разных контекстах с разными

⁵ Цит. по: Hardy L., Dixon C., Hsi S. From Data Collectors to Data Producers: Shifting Students' Relationship to Data // *Journal of the Learning Sciences*, 2020, 29(1). - С. 104–126.

⁶ Цит. по (22).

намерениями, что делает проблематичным саму возможность анализа датасета как неизменного во времени текста⁷.

Тем не менее, ряд авторов рассматривает сферу науки о данных и «дата-грамотность» в русле критической теории. Этот подход в образовательных инициативах подкрепляется «дискурсом продуктивности», сфокусированном не на углублении технических навыков анализа данных, а на массовом развитии «критической дата-грамотности» (critical data literacy) для общественного блага и расширения гражданских прав и возможностей (31). Практики и исследователи этого направления часто обращаются к открытым правительственным и международным данным, касающимся широкого круга вопросов высокой общественной значимости (32). Примерами инициатив, направленных на повышение дата-грамотности различных общественных групп являются программы по журналистике данных ((33); (34); (35)), воркшопы, адресованные некоммерческим организациям и общественным активистам ((22); (36)).

Однако, несмотря на риторику прозрачности, окружающую публично доступные государственные датасеты, многие специалисты и за рубежом, и в России обращают внимание на противоречивые мотивы, сопровождающие их публикацию ((37); (38)).⁸

Спектр компетенций и образовательных программ в области науки о данных

В настоящее время в литературе описано нет так много инициатив по преподаванию науки о данных учащимся подросткового и юношеского возраста; некоторые из них, позволяющие судить о разбросе содержания, длительности и целевой аудитории, приведены в Таблица 2. В то же время существует большое количество программ профессионального образования в сфере науки о данных бакалаврского и магистерского уровня ((39); (40); (41)), а соответствующих онлайн-курсов, отвечающих растущим запросам кадрового рынка в этой сфере, еще больше, и

⁷ Это не касается случаев, когда собственно текст как таковой является датасетом, обрабатываемым и исследуемым машинными средствами обработки естественного языка.

⁸ См. также: (1) Линделл Д. Раскрытие данных о заболеваниях россиян [Электронный ресурс] / Инфокультура. Опубликовано 15.03.2020. URL <https://youtu.be/fepraC4rfr4> ; (2) Открытость и открытые данные в современной России. Дискуссия [Электронный ресурс] / Инфокультура. Опубликовано 20.03.2020. URL <https://youtu.be/-oXQTTEsAN8>

их предметное содержание общедоступно⁹. Однако, как отмечают Heinemann и др. (42), программы высшего образования не применимы на уровне школы т.к. они слишком сфокусированы на углубленном изучении технической стороны анализа данных и уделяют недостаточно внимания рассмотрению вопросов влияния данных на жизнь общества и индивида, важных для подростков. Структура профессиональных курсов связана с производственным циклом стандартного проекта исследования данных, а развиваемые в них компетенции хорошо описаны (компьютерное и статистическое мышление; математические основы, алгоритмы, построение и оценка моделей машинного обучения, подготовка данных, представление результатов и т.д. (39). Heinemann и др. (42) резонно полагают, что копировать всю модель CRISP-DM для программы школьного уровня не имеет смысла, т.к. учащиеся не находятся в реальном производственном контексте, а, стало быть, такие фазы как «понимание бизнес-требований» и «внедрение» нужно опустить или адаптировать к учебному процессу, как это, например сделано у Srikant и Aggarwal (43), хотя, в целом, цикл CRISP-DM задает вполне работающую рамку для образовательной программы и для реализации коллективных учебных проектов.

Приведение этой модели к менее абстрактным и более простым для учащихся этапам делается через фокус на более низкоуровневых практических навыках работы с данными в той или иной программной среде (например, с использованием языков Python или R). Так, Erickson и др. (44) предлагают продвигаться по фазам исследования данных, делая элементарные «шаги» (data moves) – выполняя конкретные действия с данными, без которых не обходится ни один дата-проект, например, фильтрация, группировка данных или построение диаграмм. По мнению авторов и цитирующих их специалистов, этот «ремесленный» уровень работы с данными готовит школьников к лучшему пониманию более сложного материала - индуктивной статистики и математических основ машинного обучения.

Приведем в качестве примера некоторые “шаги” (data moves) из программы интенсива для старшеклассников “Дата Кампус: основы анализа данных и машинного обучения”¹⁰: импорт-экспорт данных, проверка и исправление типов данных, кодирование категориальных переменных, переименование-добавление-удаление

⁹ Например, на сайтах Coursera <https://www.coursera.org/browse/data-science> и edX <https://www.edx.org/course/subject/data-science>.

¹⁰ Дерябин А.А. Дата Кампус: игры с большими данными [Электронный ресурс] // EdExpert. Эффективные решения в образовании. URL <http://edexpert.ru/datacampus> (дата обращения 10.04.2020).

столбцов, устранение или заполнение пустых значений, поиск выбросов данных, вывод описательных статистик, вывод информации о датасете, построение диаграмм рассеивания, обращение к произвольным элементам и фильтрация датасета.

Таблица 2. Примеры учебных модулей науки о данных для учащихся школьного возраста.

Источник	Содержание модуля/курса	Аудитория, формат
Heinemann и др., (2018). Drafting a Data Science Curriculum for Secondary Schools (42)	1. Статистическое мышление и методы статистики 2. Большие данные и искусственный интеллект (деревья решений и нейросети) 3. Учебные проекты, основанные на реальных данных. 4. Наука о данных и общество	Германия; школьники 15–18 лет Учебный курс – часть обязательной школьной программы, 18 учащихся.
Srikant, Aggarwal (2017). Introducing Data Science to School Kids (43)	Полный цикл машинного обучения с учителем: подготовка данных, визуализация, построение и тестирование модели.	Индия, США; школьники 10-15 лет Одно занятие на половину дня, всего 71 учащихся в 4 городах.
Bryant (2019). A Middle-School Camp Emphasizing Data Science and Computing for Social Good (45)	Основы статистики, парное программирование (BlockPy, BBC micro:bits, Jupyter Notebooks), синтаксис Python, методы работы с данными и визуализацией; проекты: «наука о данных для социального блага».	США, средняя школа Летний лагерь, 1 неделя.
Dryer (2018). A Middle-School Module for Introducing Data-Mining, Big-Data, Ethics and Privacy Using RapidMiner and a Hollywood Theme (46)	Добыча данных, предиктивная аналитика, визуализация, большие данные, этика и приватность, использование ПО RapidMiner, датасет Titanic, построение рекомендательной системы про кинофильмы.	США, школьники 11-19 лет (большинство 12-14) 90-минутный воркшоп
Datta & Nagabandi (2018). Integrating Data Science and R Programming at an Early Stage (47)	Две части: (1) введение в науку о данных через игры в сбор данных, чистку, визуализацию. Базовая информация о машинном обучении (классификация, кластеризация)	США, школьники 11 лет Воркшоп 15 дней по 2 часа; 100 учащихся.

	через групповые активности без использования компьютера; (2) введение в программирование на языке R.	
Mariescu-Istodor (2019). Machine Learning Exercises for High School Students (48)	Написание кода для обработки изображений и передачи в готовую модель машинного обучения. Классификация изображений. План урока: введение в обработку изображений (15 мин), описание метода (20 мин), групповая работа (30 мин), работа с приложением машинного обучения (15 мин), эксперименты (20 мин).	Румыния, школьники 13 и 19 лет. 2-часовой урок
Wolff и др. (2019). Exploring design principles for data literacy activities to support children's inquiry from complex data (49)	Полевые исследования с сбором и анализом данных на тему «умный город» (использование солнечной энергии и потребление электричества).	Великобритания, дети 10-14 лет. 4 полевых исследования длительностью несколько недель. Всего 67 учащихся.

Какие данные использовать для обучения?

С точки зрения ряда специалистов, наибольший учебный эффект на дата-грамотность, критическое мышление и предметные знания старшеклассников имеет погружение в исследование аутентичных научных данных (50). Авторы широко цитируемого подхода «дата-шагов» Erickson и др. считают, что вводный курс науки о данных должен давать учащимся задачи с разнообразными по содержанию, необработанными, «непослушными» данными (44). Wolff и др. (49), работая с детьми 10-14 лет, также полагают, что следует предъявлять учащимся не упрощенные, но комплексные данные. Однако «комплексные» данные не обязательно означает «большие», а «реальные» не обязательно означает «подлинные». Чтобы избежать путаницы в терминах, приводим ниже адресованный преподавателям разбор характеристик аутентичных учебных данных, обобщенный в Таблица 3.

Под аутентичными, или подлинными, данными понимаются «достоверные количественные и качественные данные, извлеченные из феноменов реальной жизни». Наоборот, неаутентичные данные могут быть искусственно сгенерированы в демонстрационных целях или могут быть результатом манипулирования данными для

достижения задуманного учебного результата или интерпретации. Аутентичные данные могут собраны с использованием различных измерительных методов и инструментов (например, данные сенсоров), сгенерированы компьютерными моделями и симуляциями или получены из онлайн-репозитория и научных публикаций ((51); (52); (53)). Kjelvik и Schultheis (50) имеют в виду в первую очередь естественнонаучные данные, однако с некоторыми допущениями и в зависимости от содержания учебной программы преподаватель вполне может отнести к ним неадаптированные экономические, социологические и проч. данные. Сбор, анализ, интерпретация аутентичных датасетов потенциально помогают учащимся улучшить навыки критического мышления, углубить предметные знания и связать свои количественные аналитические навыки с сферой науки о данных ((54); (55)).¹¹ Ниже представлены характеристики сложности учебных данных, которые могут варьироваться в зависимости от задач учебной программы и подготовленности аудитории.

Охват (scope) данных определяется числом переменных и числом значений, которые принимает каждая переменная. Простые датасеты состоят из небольшого числа переменных и содержат только ту информацию, которая нужна для ответа на вопросы исследования. Сложные, с широким охватом датасеты могут содержать много переменных, как нужных, так и не нужных для решения поставленной исследовательской задачи. «Узкие» датасеты хороши для наглядной демонстрации зависимых и независимых переменных и отношений между ними; работа с «широкими» датасетами может представлять собой дополнительный вызов для учащихся, которые должны определить, какие переменные им нужны, а какие нет, и может стимулировать их исследовательскую активность, которая может вести к творческой переформулировке поставленных задач и выявлению непредвиденных связей между переменными.

Селективность (selection) касается принятия решений о том, какие переменные нужно определить, какие данные, откуда и при помощи каких инструментов необходимо собрать для решения поставленной исследовательской задачи. В простейшем случае переменные задаются преподавателем. В случае, если учащимся предлагается сделать это самостоятельно, им, тем самым, предоставляется возможность самим определить свою образовательную траекторию (56). Последнее,

¹¹ Цит.по: Kjelvik & Schultheis (50).

разумеется, обеспечивает более богатый образовательный и исследовательский опыт (5); (55); (57)).

Курирование (curation) данных – это организация и интеграция данных, собранных из разных источников, подготовка их к анализу и обеспечение их целостности в этом процессе; курирование может включать в себя обнаружение и сбор данных, аннотирование, подготовку и очистку, представление, организацию доступа к данным и их хранение. Преподаватель зачастую обеспечивает учащихся хорошо организованными, готовыми к использованию датасетами, чтобы облегчить их задачу. Однако, когда курирование делегируется студентам, у них появляется возможность освоить новые навыки работы с данными (5): синтез нескольких датасетов в один предполагает приведение их к единому формату и размерности, кодирования категориальных переменных в численные и проч., что требует умений работы с электронными таблицами и соответствующими программными методами.

Беспорядок (messiness) – наличие в датасете пропусков и выбросов, случайных и систематических ошибок, влияющих на изменчивость переменных. Подготовка сырых данных – обычно едва ли не самый длительный этап в цикле промышленного дата-проекта, и навыки чистки «грязных» данных имеют большую ценность как для лучшего понимания учащимися источников изменчивости признаков в датасете, так и для приобретения практических навыков манипулирования данными.

Размер (size) данных определяется числом переменных (число колонок в таблице данных) и числом значений переменной (число строк). В учебном процессе размер учебного датасета может варьироваться от маленького на начальных этапах обучения, до большого в ходе выполнения студенческого проекта.

Таблица 3. Характеристики сложности учебных данных по Kjelvik & Schultheis (50).

Характеристика	← Простота		Сложность →
Охват (scope)	Узкий: есть только нужные данные		Широкий: есть нужные и ненужные данные
Селективность (selection)	Переменные заданы преподавателем	Студенты выбирают переменные из предложенного датасета	Студенты самостоятельно собирают датасет с нужными им переменными

Курирование (curation)	Полное: датасет дается студентам в готовом к анализу виде	Частичное: сырые данные, готовые к анализу, но не аннотированные	Отсутствует: студенты должны аннотировать сырые данные и подготовить их к анализу, трансформировать	Синтез: студенты должны объединить несколько датасетов и подготовить их к анализу
Размер (size)	Маленький: несколько переменных (столбцов) и строк		Большой: много переменных (столбцов) или строк	
Беспорядок (messiness)	Чистые: нет пропусков данных и выбросов, низкая изменчивость		Грязные: есть пустые значения и выбросы, высокая изменчивость	

Обеспечение учащихся опытом работы с разными типами данных вместе с обсуждением контекста использования этих данных и того, как именно эти данные помогают ответить на вопросы их исследования может быть важным для развития дата-грамотности (58). Ряд педагогических инициатив позволяет предположить, что взаимодействие с аутентичными данными положительно влияет на дата-грамотность учащихся ((56); (59)). Вместе с тем, не многие исследования проверяли эффективность учебных материалов, использующих аутентичные данные (60), а исследований, сравнивающих эффекты аутентичных и неаутентичных данных на результаты обучения в настоящее время нами не обнаружено. Чтобы продемонстрировать эту зависимость требуются исследования, которые бы изолировали воздействие аутентичных данных на дата-грамотность, интересы учащихся и их вовлеченность в изучение данных через манипулирование использованием аутентичных и неаутентичных данных в процессе обучения.

Релевантность данных интересам учащихся как условие эффективной реализации образовательной задачи

Для эффективного обучения данные и результаты работы с ними должны быть «присвоены» студентами, а для этого они должны быть не только аутентичными, но и релевантными интересам учащихся. Когда учащиеся не представляют, как данные, с которыми им предложили работать, связаны с реальной жизнью, учебные задачи и

действия (обращаются к базе данных, работают с датасетом, интерпретируют диаграммы) могут им представляться затруднительными или бессмысленными (50). Некоторые могут испытывать трудности с интерпретацией неаутентичных данных (9); (61)) или с применением новых умений в других контекстах (62).

Напротив, аутентичные данные, связанные с предметным содержанием, с понятным контекстом их создания и в ясной связи с проблемами реального мира потенциально интересней и привлекательней для учащихся ((58); (63)), и от них поступает обратная связь, которая говорит об эмоциональной вовлеченности в работу с такими данными ((55); (49)).

Как обеспечить это присвоение данных в образовательном процессе? Через релевантную интересам учеников тематику учебных задач, сопутствующих им данным и взаимодействие с ними. Образовательная программа должна предусматривать решение таких образовательных задач, которые, с одной стороны, имеют статус «настоящих» (то есть актуальных и практически значимых) для всех ее участников, а с другой стороны, задают некоторую неопределенность как в ее интерпретации, так и в сценариях их решения¹², что может быть привнесено в образовательный процесс за счет использования необработанных аутентичных данных. Такая специфика программы позволяет обеспечить продуктивную деятельность учащихся и удерживать высокую мотивацию учащихся на использование данных и техники работы с ними как образовательного ресурса.

Wilkerson & Polman (64) отмечают, что изучение науки о данных – это не просто утилитарный процесс предметного обучения, но выстраивание многослойных отношений учащегося по отношению к данным как социокультурному тексту: каждый учащийся может обнаружить данные, которые рассказывают о нем что-то важное, будь то данные от сенсоров носимых им «умных часов», геометки в истории его навигатора или обобщенная социально-экономическая статистика о регионе его проживания.

Lee & Dubovi (65) показывают, как возникает глубокая и сложная связь с данными у семей с диабетом 1 типа в процессе использования ими устройств мониторинга и постоянных измерений. Kahn (66) описывает конструирование «семейных геобиографий» на основе данных и демонстрирует, как их интерпретации в сочетании с семейными историями миграции формируют вокруг них многослойный социоисторический контекст, влияющий на молодое поколение.

¹² Глухов П.П. Компетентностные испытания как современная форма оценки образовательных достижений // Философия образования, 2016, № 4 (67). С. 99-110.

Кроме того, для создания условий эффективной профессиональной пробы важно дать учащимся попробовать себя в разных позициях в ходе решения задач по отношению к данным – например, в разных ролях в соответствии с разделением труда внутри дата-проекта (программист, дата-инженер, менеджер проекта и предметный эксперт). Так, Wise (31), описывая свой проект, в котором студенты собирают данные о собственном обучении, выделяет три взаимодополняющих позиции по отношению к данным, между которыми переключается ее студенты в ходе учебного проекта: «внутренняя» (респондент, источник данных), «внешняя» (реципиенты, которым адресован аналитический отчет о проделанной работе, например «заказчик») и «аналитик данных» (тот, кто работает с данными).

Hardy и др. (15) подчеркивают важность становления субъектности учащихся в генерации знания через культивирование активной позиции «дата-продюсера» в противоположность традиционным, дисциплинарным образовательным подходам, нацеленным на воспроизводство заранее известного и дидактически «правильного» результата и «правильных» данных, которые словно бы ждут, чтобы их собрали. Wise (31 pp. 13-14), рассуждая о том, что должна делать образовательная программа для формирования дата-компетенций учащихся, обобщает несколько пунктов.

1. Вырабатывать критическое отношение к данным. Нужно развивать у учащихся понимание того, как человеческие, материальные и технические (от себя добавим – и институциональные) аспекты производства данных (кто сгенерировал эти данные? Как? Для чего?) взаимодействуют между собой и отражаются на конкретном датасете, особенно в обстоятельствах, когда студенты не участвуют в его генерации; что конкретные данные – это всегда лишь одна из возможных версий данных, которые могли бы существовать; какие различия могут быть у альтернативных версий и почему. Критическое отношение к происхождению данных - важная часть интерпретации данных и дата-образования.

2. Давать возможность студентам испытать сложность работы с большими, комплексными, беспорядочными датасетами. Подготовка данных, наряду с визуализацией и построением моделей машинного обучения – одна из основных задач, решаемых в реальных дата-проектах. Дать студенту ощутить «сопротивление материала» принципиально важно для понимания производственных проблем, возможностей и ограничений инструментов, применяемых для их решения.

3. Помогать студентам искать паттерны и закономерности в данных, быть способными делать выводы не только относительно отдельных наблюдений или

переменных в датасете, но делать обобщения, уметь менять масштаб рассмотрения проблемы (например, от муниципального до глобального), видеть закономерности в агрегированных данных.

4. Помогать студентам исследовать, как разные предназначения одних и тех же данных влияют на их представление и трансформацию. В общем случае, работая с данными, студенты проходят только одной «тропой», заданной преподавателем или избранной ими самими. Возврат к началу исследования того же самого датасета с новыми вопросами дает им представление о цикличности работы с данными и демонстрирует, как переопределение задачи реорганизует используемые ими инструменты, техники и ресурсы.

Очевидно, форматы и методы преподавания науки о данных связаны с тем, как образовательная программа или учебный курс определяет свою задачу: знакомство учащихся с наукой о данных как первый шаг на пути их профессионального становления в этой области или обучение навыкам дата-грамотности как одной из базовых компетенций, необходимых для жизни в современном «датафицированном» обществе вне зависимости от карьерного выбора учащегося.

Говоря о профориентационной компоненте преподавания науки о данных в подростковом и юношеском возрасте, следует учитывать актуальные требования, предъявляемые к современному специалисту: не только соотносить конкретную задачу с перечнем испытанных вариантов решения и выбирать из этого перечня оптимальный вариант, но и конструировать новые способы решения, оптимальные именно для данного конкретного случая; оформлять проблемную ситуацию и ставить задачи, исходя из неё; намечать необходимый результат для своих действий в рамках конкретного проекта или технологического процесса, и выстраивать программу действий по достижению этого результата. Для этого у специалиста должен быть сформирован тип мышления, соответствующий выбранной им сфере профессиональной деятельности, позволяющий понимать присущие этой сфере технологии, приёмы работы и тип результата.

Так, Суосо и др. (67) пишут о важности формирования “ментальной привычки” обращения к данным как способа мышления, вопрошания, и решения задач. Finzer (14) также приверженец идеи, что в обучении в школе нужно прививать специфическое мышление, взгляд на мир сквозь призму данных. Это видение прослеживается и у авторов, пишущих о разработке образовательных программ по Data Science и статистике на уровне высшей школы ((39); (68)).

Уровни компетенций в области дата-грамотности

Поскольку исследование или анализ – это базовый тип деятельности в области науки о данных, необходимо рассмотреть вопрос оценки аналитических (исследовательских) компетенций учащихся.

Необходимо различать продуктивную и репродуктивную деятельность и, соответственно, продуктивные и репродуктивные компетенции.

Репродуктивная деятельность представляет собой действия, выполняемые по ранее усвоенному алгоритму, и репродуктивные компетенции связаны с освоением этого алгоритма, пониманием условий его применения и способностью воспользоваться им. Репродуктивные компетенции оцениваются посредством выполнения заданий, имеющих решение.

Продуктивная деятельность – это деятельность без *заранее известного алгоритма или образца*, в заведомо уникальной ситуации, а также деятельность по созданию нового образца. Формы оценки продуктивных компетенций связаны с открытыми, проблемными ситуациями, которые необходимо еще превратить в задачу. То есть это ситуации, само понимание которых требует применения знаний и тех или иных способов деятельности.¹³

Репродуктивная деятельность, применительно к практике анализа данных, заключается в применении стандартных методов и процедур, составляющих технические навыки работы с данными, например:

- импорт-экспорт данных,
- проверка и исправление типов данных,
- кодирование категориальных переменных,
- переименование-добавление-удаление признаков,
- устранение или заполнение пустых значений,
- поиск выбросов данных,
- вывод описательных статистик, вывод информации о датасете,
- построение диаграмм рассеивания,
- обращение к произвольным элементам и фильтрация датасета,
- разведочный анализ данных.

¹³ Глухов П.П. Компетентностные испытания как современная форма оценки образовательных достижений // Философия образования, 2016, № 4 (67). С. 99-110.

Даже создание и настройку модели машинного обучения учениками, не знакомыми с математическими основаниями алгоритмов, стоящими за нею, можно рассматривать как репродуктивный навык. Опыт показывает, что учащийся, владеющий программированием, вполне может скопировать код образца и минимально отредактировать его под свою задачу. Далее он в состоянии методом проб и ошибок подобрать параметры модели, не вникая в их математический смысл, чтобы она выдавала приемлемое значение метрики качества. Более того, пользуясь шаблонным кодом, он(а) может передать одни и те же данные сразу нескольким моделям с настройками «по умолчанию», использующим разные алгоритмы (например, дерево решений, градиентный бустинг, нейронная сеть) и выбрать в качестве решения ту, которая дает наилучший результат. В данном случае мы скорее имеем дело с хорошими репродуктивными навыками, но не с продуктивной деятельностью.

Как было сказано выше, становление субъектности учащихся в генерации знания через аналитические практики происходит через культивирование их активной позиции «дата-продюсера» в противоположность традиционным, дисциплинарным образовательным подходам, нацеленным на воспроизводство заранее известного и дидактически «правильного» результата и «правильных» данных (69), в осуществлении перехода от репродуктивной к продуктивной деятельности.

Это требует от учащихся владения компетенциями, задающимися соответствующими элементами процесса по исследованию данных (CRISP-DM), который включает в себя следующие фазы: *понимание целей – начальное изучение данных – подготовка данных – моделирование – оценка – внедрение* (2).

В то время как для подлинно продуктивной деятельности на фазах «моделирования» и «оценки» учащимся действительно необходимы знания математической статистики, теорий вероятностей и алгебры, выходящие за рамки школьного курса, фазы «понимания», «начального изучения данных» и «подготовки данных» вполне могут быть реализованы в продуктивном ключе.

Этот уровень компетенций включает в себя такие компоненты:

— владение схемами и принципами анализа, выражающегося в способности анализировать многомерные объекты (напр., региональное или государственное управление, производственная ситуация) с точки зрения построения гипотез, постановки аналитических задач, интерпретации результатов.

- критическое отношение к данным; понимание того, как человеческие, материальные, технические, институциональные аспекты производства данных взаимодействуют между собой, отражаются на конкретном наборе данных и их презентации.

– понимание цикличного характера работы с данными, способность скорректировать аналитическую задачу в зависимости от характера имеющихся данных, понимание того, как переопределение задачи реорганизует используемые инструменты, техники и ресурсы.

– манипулирование данными и их визуализация с целью донести до целевой аудитории результаты анализа в соответствии с поставленной задачей.

Опыт авторов настоящей работы по реализации программы обучения школьников подросткового и младшего юношеского возраста основам анализа данных и машинного обучения показывает, что именно на начальном этапе решения учебной задачи или разработки учебного аналитического проекта, включающего в себя «понимание» задачи или ее постановку, «начальное изучение данных», их поиск и «подготовку» под задачу, учащиеся испытывают наибольшие сложности.

Эти сложности в меньшей степени выражены, когда объектом исследования или дата-проекта является какая-либо инженерная система (например, беспилотный автомобиль), и в большей степени, когда объект или предмет – социокультурная сущность (например, «регион» или «счастье»). В первом случае учащимся проще определить цели, задачи своего проекта, понять какие данные им необходимы для их решения. Во втором случае уже на этапе формулировки гипотез учащиеся сталкиваются с тем, что объект трудно «схватить», концептуализировать, дать ему определение, выделить его существенные признаки и их корреляты из разных дисциплинарных областей и подобрать соответствующие данные.

Формирование образа будущей профессии дата-аналитика и связанного с ней типа мышления может обеспечиваться практико-ориентированным образованием, где в качестве единицы при построении учебного процесса используется реально существующая проблема, которая переживается учеником как значимая.¹⁴ Эта проблема может быть описана посредством количественных данных и оформлена в виде образовательной задачи, организующей одновременно пробное практическое

¹⁴ Глухов П.П. Практическое мышление как компетентностный образовательный результат / Социальные коммуникации и эволюция обществ. Материалы V Международной научно-практической конференции. Под редакцией И.А. Вальдмана; Новосибирск: 2016. С. 447-453.

действие ученика и его рефлексию в отношении используемых им инструментов анализа данных.

Поскольку в современных формах инженерно-математического образования все чаще звучит требование к студенту уже на старте получения профессионального образования иметь образ будущей профессии, работа по формированию у будущих дата-аналитиков специфически профессионального мышления и основ профессиональной культуры должна происходить ещё в период их обучения в школе, в режиме пропедевтики будущей деятельности.

Список литературы

- 1] Ward J.S., Baker A., Undefined By Data: A Survey of Big Data Definitions. [Электронный ресурс]. // Cornell University Library, 20 Sept. 2013. URL <http://arxiv.org/abs/1309.5821> (дата обращения 10.04.2020), 2013.
- 2] Chapman P. и др., CRISP-DM 1.0: Step-by-step data mining guide, / CRISP-DM Consortium, 2000. [Электронный ресурс] URL <https://www.the-modeling-agency.com/crisp-dm.pdf> (дата обращения 10.04.2020), 2000.
- 3] T. Matzner, "Data science education as contribution to media ethics," in // *Paderborn Symposium on Data Science Education 2017: The Collected Extended Abstracts*, Paderborn, University of Paderborn, 2018.
- 4] M. Schield, "Information literacy, statistical literacy and data literacy," // *IASSIST Quarterly*, vol. 28, no. 2, p. 6–11, 2018.
- 5] Carlson J. R. и др., "Determining data information literacy needs: A study of students and research faculty. Paper 23," // *Libraries Faculty and Staff Scholarship and Research*, p. [Электронный ресурс] URL https://docs.lib.purdue.edu/lib_fsdocs/23/ (дата обращения 10.04.2020), 2011.
- 6] Mandinach E. B., Gummer E. S., "A systemic view of implementing data literacy in educator preparation," *Educational Researcher*, vol. 42, no. 1, p. 30–37, 2013.
- 7] Gibson J. P., Mourad T. , "The growing importance of data literacy in life science education," *American Journal of Botany*, vol. 105, no. 12, p. 1–4, 2018.
- 8] Steen L. A., *Achieving quantitative literacy: An urgent challenge for higher education*, Washington, DC: Mathematical Association of America, 2004.
- 9] Piatek-Jimenez K. и др., "Helping students become quantitatively literate," / *Mathematics Teacher*, vol. 105, no. 9, p. 692–696.
- 10] Boersma S., Klyve D., "Measuring habits of mind: Toward a prompt-less instrument for assessing quantitative literacy [Электронный ресурс]," // *Numeracy: Advancing Education in Quantitative Literacy*, vol. 6, no. 1, p. ar6. URL <https://scholarcommons.usf.edu/numeracy/vol6/iss1/art6/> (дата обращения 10.04.2020), 2013.

- 11] Vacher H. L., "Looking at the Multiple Meanings of Numeracy, Quantitative Literacy, and Quantitative Reasoning. [Электронный ресурс]," // *Numeracy*, vol. 7, no. 2, p. Article 1. URL <https://scholarcommons.usf.edu/numeracy/vol7/iss2/art1/> (дата обращения 10.04.2020), 2014.
- 12] Mayes R. L. и др., "Quantitative reasoning in environmental science: A learning progression," // *International Journal of Science Education*, vol. 36, no. 4, p. 635–658, 2014.
- 13] D. Conway, "The data science venn diagram. [Электронный ресурс]," pp. URL <http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram> (дата обращения 10.04.2020), 2010.
- 14] W. Finzer, "The data science education dilemma [Электронный ресурс]," // *Technology Innovations in Statistics Education*, vol. 7, no. 2, p. URL <https://escholarship.org/uc/item/7gv0q9dc> (дата обращения 10.04.2020), 2013.
- 15] L. D. C. H. S. Hardy, "From Data Collectors to Data Producers: Shifting Students' Relationship to Data," // *Journal of the Learning Sciences*, vol. 29, no. 1, p. 104–126., 2020.
- 16] D. Ihde, "Heidegger's philosophy of technology," // *Technics and Praxis. Boston studies in the philosophy of science*, vol. 24, p. 103–129, 1979.
- 17] P. Kroes, "Physics, experiments, and the concept of nature," in *The Philosophy of scientific experimentation*, Pittsburgh, PA, University of Pittsburgh Press, 2003, p. 68–86.
- 18] S. Tala, "Unified view of science and technology for education: Technoscience and technoscience education," // *Science & Education*, vol. 18, no. 3–4, p. 275–298, 2009.
- 19] B. Latour, *Science in action: How to follow scientists and engineers through society*, Harvard University Press, 1987.
- 20] Machamer P., Osbeck L., "The social in the epistemic," in *Science values and objectivity*, University of Pittsburgh Press, 2004, p. 78–89.
- 21] Pinch T. J., Bijker W. E., "The social construction of facts and artefacts: Or How the sociology of science and the sociology of technology might benefit each other," // *Social Studies of Science*, vol. 14, no. 3, p. 399–441, 1984.
- 22] A. Fotopoulou, "Conceptualising critical data literacies for civil society organisations: agency, care, and social responsibility dilemma," // *Information Communication and Society*, p. doi:10.1080/1369118X.2020.1716041, 2020.
- 23] Kellner D., Share J., "Toward critical media literacy: Core concepts, debates, organizations, and policy," // *Discourse: Studies in the Cultural Politics of Education*, vol. 26, no. 3, p. 369–386, 2005.
- 24] S. Livingstone, "Media literacy and the challenge of new information and communication technologies," // *The communication review*, vol. 7, no. 1, pp. 3-14, 2004.
- 25] Luke A., Freebody P., "Shaping the social practices of reading," // *Constructing Critical Literacies: Teaching and Learning Textual Practice*, vol. 6, p. 460–475, 1997.

- 26] K. Crawford, "The hidden biases in big data [Электронный ресурс]," // *Harvard Business Review Blog Network*, pp. URL <http://blogs.hbr.org/2013/04/the-hidden-biases-in-big-data/> (дата обращения 10.04.2020), 2003.
- 27] G. Bowker, "Data flakes: An afterword to 'raw data' is an oxymoron," in *'Raw data' is an oxymoron*, Cambridge, MA, MIT Press, 2013, p. 167–171.
- 28] Boyd D., & Crawford K., "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon," // *Information, Communication & Society*, vol. 15, no. 5, p. 662–679, 2012.
- 29] P. Dourish, "Algorithms and their others: Algorithmic culture in context," // *Big Data & Society*, vol. 3, no. 2, p. doi:10.1177/2053951716665128, 2016.
- 30] Pangrazio L., Sefton-Green J., "The social utility of 'data literacy'," // *Learning, Media and Technology*, p. doi:10.1080/17439884.2020.1707223, 2019.
- 31] A. Wise, "Educating Data Scientists and Data Literate Citizens for a New Generation of Data," // *Journal of the Learning Sciences*, vol. 29, no. 1, p. 165–181, 2020.
- 32] Pangrazio L., Selwyn N., "'Personal data literacies': A critical literacies approach to enhancing understandings of personal digital data," // *New Media and Society*, vol. 21, no. 2, p. 419–437, 2019.
- 33] J. Hewett, "Learning to teach data journalism: Innovation, influence and constraints," // *Journalism*, vol. 17, no. 1, p. 119–137, 2016.
- 34] Burns L. S., Matthews B. J., "First Things First: Teaching Data Journalism as a Core Skill," // *Asia Pacific Media Educator*, vol. 28, no. 1, p. 91–105, 2018.
- 35] C. Graham, "A DIY, Project-based Approach to Teaching Data Journalism," // *Asia Pacific Media Educator*, vol. 28, no. 1, p. 67–77, 2018.
- 36] Carroll J. M. и др., "Strengthening Community Data: Towards Pervasive Participation," in / *Proceedings of the 19th Annual International Conference on Digital Government Research, May 30-June 1, 2018, Delft, Netherlands*, New York, 2019.
- 37] Halford S., Pope C., Weal M., "Digital futures? Sociological challenges and opportunities in the emergent semantic web," // *Sociology*, vol. 47, no. 1, p. 173–189, 2013.
- 38] Langlois G. и др., "Networked publics: the double articulation of code and politics on Facebook," // *Canadian Journal of Communication*, vol. 34, no. 3, p. 415–434, 2009.
- 39] De Veaux, R. и др., "Curriculum guidelines for undergraduate programs in data science," // *Annual Review of Statistics and Its Application*, vol. 4, no. 1, p. 15–30, 2017.
- 40] B. Baumer, "A data science course for undergraduates: Thinking with data," // *The American Statistician*, vol. 69, p. 334–342, 2015.
- Anderson P. и др., "An undergraduate degree in data science: Curriculum and a decade of implementation experience," in / *SIGCSE 2014 - Proceedings of the 45th ACM*

- 41] *Technical Symposium on Computer Science Education*, 2014.
- Heinemann B. и др., «Drafting a data science curriculum for secondary schools,» в /
- 42] *ACM International Conference Proceeding Series*, 2018.
- Srikant, S., & Aggarwal, V., "Introducing data science to school kids," in / *Proceedings of the Conference on Integrating Technology into Computer Science Education, ITiCSE*, 2017.
- 43]
- Erickson T. и др., "Data Moves: one key to data science at school level," in /
- 44] *Proceedings of the International Conference on Teaching Statistics (ICOTS-10)*, 2018.
- Bryant C. и др., "A middle-school camp emphasizing data science and computing for social good," in // *SIGCSE 2019 - Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, 2019.
- 45]
- Dryer A., Walia N., Chattopadhyay A., "A middle-school module for introducing data-mining, big-data, ethics and privacy using rapidminer and a Hollywood theme," in / *SIGCSE 2018 - Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, 2018.
- 46]
- Datta S., Nagabandi V., "Integrating data science and R programming at an early stage," in / *IEEE 4th International Conference on Soft Computing and Machine Intelligence, ISCFMI 2017*, 2018.
- 47]
- Mariescu-Istodor R., Jormanainen I., "Machine Learning Exercises for High School Students [Электронный ресурс]," in / *Proceedings of the 19th Koli Calling International Conference on Computing Education Research*, 2019.
- 48]
- Wolff A., Wermelinger M., Petre M., "Exploring design principles for data literacy activities to support children's inquiries from complex data," // *International Journal of Human Computer Studies*, vol. 129, p. 41–54. doi:10.1016/j.ijhcs.2019.03.006, 2019.
- 49]
- Kjelvik M. K., Schultheis, E. H., "Getting messy with authentic data: Exploring the potential of using data from scientific research to support student data literacy," // *CBE Life Sciences Education*, vol. 18, no. 2, pp. 1-8, 2019.
- 50]
- Magnusson S. J. и др., "How should learning be structured in inquiry-based science instruction? Investigating the interplay of 1st- and 2nd-hand investigations," in / *In Proceedings of the 6th international conference on learning sciences*, Alpharetta, GA, 2004.
- 51]
- Hug B., McNeill K. L., «Use of first-hand and second-hand data in science: Does data type influence classroom conversations?,» в // *International Journal of Science Education*, 2008.
- 52]
- Kerlin, S. C., McDonald, S. P., & Kelly, G. J., "Complexity of secondary scientific data sources and students' argumentative discourse," // *International Journal of Science Education*, vol. 32, no. 9, p. 1207–1225, 2010.
- 53]
- Mourad T., Grant B. W., Gram W. K., "Engaging undergraduate students in ecological investigations using large, public datasets," // *Teaching Issues and Experiments in Ecology*, vol. 8, 2012.
- 54]
- Langen T. A. и др., "Using large public datasets in the undergraduate ecology classroom," // *Frontiers in Ecology and the Environment*, vol. 12, p. 362–363, 2014.
- 55]
- Gould R., Sunbury S., Dussault M., "In praise of messy data," // *Science Teacher*, vol.

- 56] 81, no. 1, p. 31, 2014.
- Calzada Prado, J., Marzal, M., "Incorporating data literacy into information literacy programs: Core competencies and contents," // *Libri*, vol. 63, no. 2, p. 123–134, 2013.
- 57]
- Schultheis E. H., Kjolvik, M. K., "Data Nuggets: Bringing real data into the classroom to unearth students' quantitative and inquiry skills," // *American Biology Teacher*, vol. 77, no. 1, pp. 19-29, 2015.
- 58]
- Kastens K. A., Krumhansel R., Baker I., "Thinking big—Transitioning your students from working with small, student-collected data sets towards “big data”," // *Science Teacher*, vol. 82, no. 5, p. 25–31, 2015.
- 59]
- Aikens M. L., Dolan E. L., "Teaching quantitative biology: Goals, assessments, and resources," // *Molecular Biology of the Cell*, vol. 25, p. 3478–3481, 2014.
- 60]
- CCSSI, "Common Core State Standards for Mathematics [Электронный ресурс].," in // *Common Core State Standards Initiative*, 2014.
- 61]
- E. Borges-Rey, "Data literacy and citizenship: Understanding “big data” to boost teaching and learning in science and mathematics," in *Handbook of research on driving STEM learning with educational technologies*, Hershey, PA, IGI Global, 2017, p. 65–79..
- 62]
- Hulleman C. S., Harackiewicz J. M., "Promoting interest and performance in high school science classes," // *Science*, vol. 326, no. 5958, p. 1410–1412, 2009.
- 63]
- Wilkerson M. H., Polman J. L., "Situating Data Science: Exploring How Relationships to Data Shape Learning," // *Journal of the Learning Sciences*, vol. 29, no. 1, p. 1–10. doi:10.1080/10508406.2019.1705664, 2020.
- 64]
- Lee V. R., Dubovi I., "At home with data: Family engagements with data involved in type 1 diabetes management," // *Journal of the Learning Sciences*, vol. 29, no. 1, pp. 11-31. doi:10.1080/10508406.2019.1666011, 2020.
- 65]
- J. Kahn, "Learning at the Intersection of Self and Society: The Family Geobiography as a Context for Data Science Education," // *Journal of the Learning Sciences*, vol. 29, no. 1, pp. 57-80, 2020.
- 66]
- Cuoco, A., Goldenberg, E. P., Mark, J., "Habits of Mind: an organizing principle for mathematics curriculum.," // *Journal of Mathematical Behavior*, vol. 15, no. 4, pp. 375-402, 1997.
- 67]
- Hardin, J., Hoerl, R., и др. Data Science in Statistics Curricula: Preparing Students to “Think with Data” // *The American Statistician*, p. doi: 10.1080/00031305.2015.1077729, 2015.
- 68]
- Hardy, L., Dixon, C., Hsi, S. From Data Collectors to Data Producers: Shifting Students' Relationship to Data // *Journal of the Learning Sciences*, vol. 29, no. 1, p. 104–126., 2020.
- 69]

Информация об авторах

Дерябин Андрей Александрович, MSc Social Psychology, научный сотрудник Научно-исследовательского центра социализации и персонализации образования детей Федерального института развития образования Российской академии народного хозяйства и государственной службы при Президенте Российской Федерации (Москва). Адрес: 125319, Москва, ул. Черняховского, д.9, стр.1. E-mail: andred [at] yandex.ru

Попов Александр Анатольевич, доктор философских наук, доцент, заведующий научно-исследовательским сектором "Открытое образование" научно-исследовательского центра социализации и персонализации образования детей Федерального института развития образования Российской академии народного хозяйства и государственной службы при Президенте Российской Федерации (Москва), заведующий лабораторией компетентностных практик образования Института системных проектов Московского городского педагогического университета (Москва), профессор кафедры социологии и массовых коммуникаций гуманитарного факультета Новосибирского государственного технического университета (Новосибирск). Адрес: 125319, Москва, ул. Черняховского, д.9, стр.1. E-mail: actor [at] mail.ru

Глухов Павел Павлович, научный сотрудник Научно-исследовательского центра социализации и персонализации образования детей Федерального института развития образования Российской академии народного хозяйства и государственной службы при Президенте Российской Федерации (Москва). Адрес: 125319, Москва, ул. Черняховского, д.9, стр.1. E-mail: gluhovpav.pav [at] gmail.com

АНГЛОЯЗЫЧНЫЙ БЛОК

Authors:

Andrey A. Deryabin^a

E-mail: andred [at] yandex.ru

Alexander A. Popov^a

E-mail: actor [at] mail.ru

Pavel P. Gluhov^a

gluhovpav.pav [at] gmail.com

^a Russian Presidential Academy of National Economy and Public Administration, Moscow, Russia

DATA-LITERACY AND DATA SCIENCE: EDUCATIONAL APPROACHES AND SOLUTIONS

Summary:

Currently, there are few initiatives in science learning publications on Data Science learning and curriculum development for middle and secondary schools. Following the fundamental changes in economy and labour market, the development and implementation of curricula in Data Science and Machine Learning at all levels of education become important with special emphasis on the integration of Mathematics, Natural, and Social Sciences. A systematic review of publications and an analysis of major discussions about the Data Science curriculum development for middle and secondary school is presented. The concept of 'data literacy' as a core competence in the digital society is discussed altogether with Data Science as, on one hand, a strand of STEM education, and as a sociocultural practice and empowering resource for community activism on another. Speaking of the target range of students' data literacy competences, some authors maintain that they can be related with those implied by Cross-Industry Standard Process for Data Mining. Special attention is paid to authenticity of training data, their relevance and adjustable complexity as prerequisites for effective learning. Features of authentic data complexity and their implications for data-

literacy instruction, critical thinking and learning opportunities are described. It is maintained that it's important to encourage students' agency and their engagement into research practices through facilitation of adopting 'data producer' position and solving practice-oriented educational tasks in contrast to traditional disciplinary educational approaches, where research tasks, 'proper' data, and 'proper' results are defined beforehand. A conclusion is drawn about the necessity of learning the basics of Data Science in secondary school for the development of student's general critical thinking, and their orientation toward scientific and engineering practices. The review primarily is based on the publications in Scopus indexed journals dated 2014-2020. The article was prepared as a part of the state-assigned research work of the Russian Presidential Academy of National Economy and Public Administration.

Keywords

big data; data literacy; data science; education; machine learning; artificial intelligence; school education; computer science education; curriculum development; social science