Формирование дата-грамотности в общем образовании: от отдельных навыков к основе человеческого потенциала

25.08.2020

- **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 РАНХиГС
- **Дерябин Андрей Александрович**
  MSc Social Psychology, научный сотрудник Научно-исследовательского центра социализации и персонализации образования детей Федерального института развития образования РАНХиГС

Технологии «на коленке»: от музыки к данным

1995

🎹 производство электронной музыки. Уже в 1990-х (Logic)

2000

🎛 персональное пр-во video (iMovie, 2000)

🖼 производство и дистрибуция контента (платформы LJ, WP) - 2000 блоги, упрощение публикации

2001

🎧📱 Музыка: iPod + iTunes. новая модель дистрибуции (2001). Napster (1999)

2005

📺 дистрибуция видео - Youtube - 2005; видео СМИ - vlogs - 2016

2007

📖 Книгоиздание (ePub Standard - 2007)

2005/ Соцсети, RSS-агрегация новостей

📱 2007 Эра Мобайл: iPhone 1gen

2009

👾2009: Angry Birds release
The Casual Revolution (Jesper Juul, 2010)

2015

ML, AI - 2015; облачные платформы, Big Data

🤖 чатботы + AI, более 20 платформ для этого - 2016

Наука о данных / Data Science

- Термин «наука о данных» (data science) широко используется, когда речь идет об использовании компьютерных методов для сбора, обработки, анализа и визуализации больших объемов данных.

<u>Бизнес-процессы</u>

- *The Big Data Analysis Pipeline*. (The Computing Community Consortium Big Data Whitepaper, 2012).
- *The Knowledge Discovery in Databases* (KDD Process, 1997)
- *The Cross Industry Standard Process for Data Mining* (CRISP-DM, 2000)

<u>Методы</u>

- Математическая статистика и теория вероятностей
- Программирование
- Обработка и визуализация данных

Дата-грамотность

- Дата-грамотность лежит на пересечении областей количественного мышления и науки о данных. Количественное мышление – это способность применять математические принципы, критическое мышление и ясную логику к решению каждодневных проблем [Steen, 2004; Piatek-Jimenez, 2012; Boersma & Klyve, 2013; Vacher, 2014; Mayes et al, 2014].
- Дата-грамотный 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 [Gibson & Mourad, 2018].
- Важны не только предметные знания, но понимание контекста генерации и распространения данных

Критическая дата-грамотность

- данные всегда несут в себе следы своего технического и материального происхождения [Ihde, 1979; Kroes, 2003];
- 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 [Ihde, 1979; Tala, 2009];
- 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 предметном, но также в экономическом и политическом контексте [Latour, 1987].

Данные в социокультурном контексте

- взаимодействие с данными является социокультурной практикой, и большие данные могут нести в себе скрытые мотивы их кураторов, напр., электоральные данные или данные здравоохранения*
- «критическая дата-грамотность» (critical data literacy) для общественного блага и расширения гражданских прав и возможностей

* (1) Раскрытие данных о заболеваниях россиян / Инфокультура (15.03.2020). URL https://youtu.be/fefpaC4rfr4

(2) Открытость и открытые данные в современной России / Инфокультура (20.03.2020). URL https://youtu.be/-oXQTTESaN8

Дата-грамотность vs. Наука о данных

Дата-грамотность	Наука о данных
Способность понимать и оценивать информацию о реальном мире, которая может быть получена из данных	Умение решать с помощью данных 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

Практические вопросы преподавания дата-грамотности и науки о данных

- Какие данные?
- Какой может быть образовательная программа?
- Какое ПО использовать для обучения?

Учебные данные

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	

Kjelvik, M. K., & Schultheis, E. H. (2019). Getting messy with authentic data: Exploring the potential of using data from scientific research to support student data literacy // *CBE Life Sciences Education*, *18*(2), 1–8. DOI:10.1187/cbe.18-02-0023

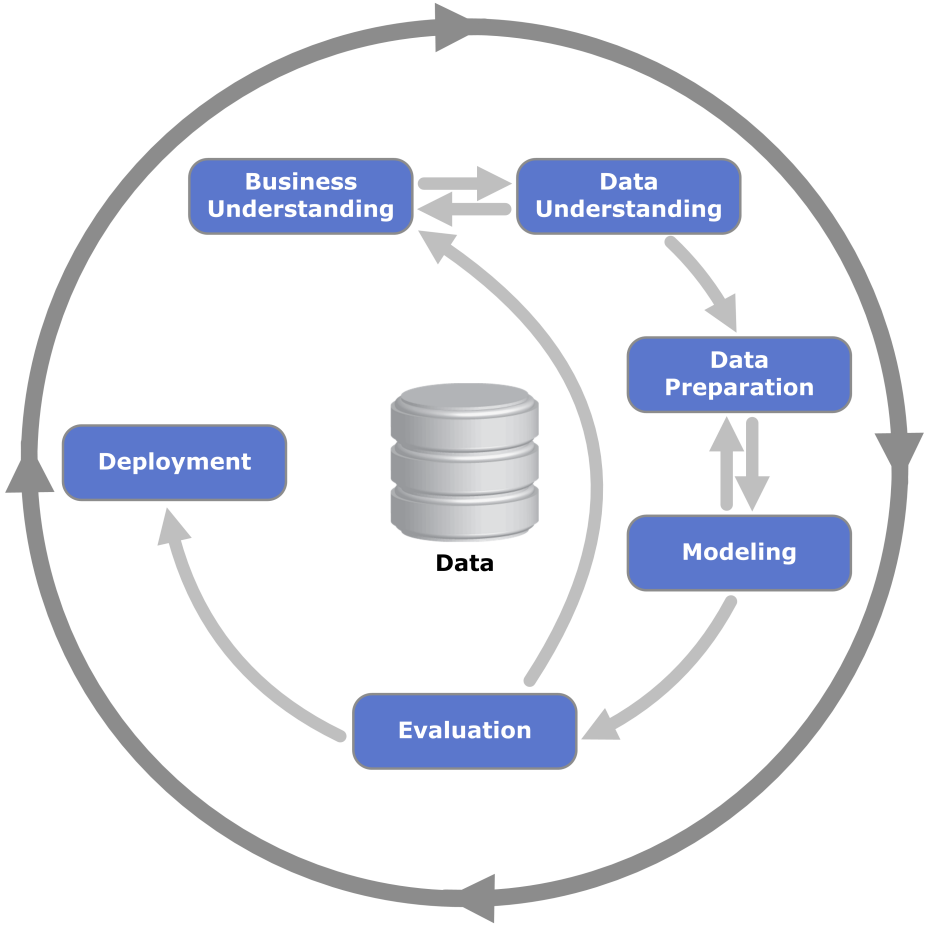
Учебные данные

- Аутентичные данные: «достоверные количественные и качественные данные, извлеченные из феноменов реальной жизни» (не искусственно созданные в дидактических целях).
- Необработанные, «сырые».
- Связанные с жизнью учащихся; реально существующая проблема (выраженная в данных), которая воспринимается учеником как значимая.
- «Присвоенные»: сгенерированы или хотя бы собраны учащимися.
- «Дата-продюсер»: учащийся как исследователь и активный создатель дата-продукта.

Фазы промышленного цикла исследования данных

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.

1. Понимание целей (*Business Understanding*)
2. Начальное изучение данных (*Data Understanding*)
3. Подготовка данных (*Data Preparation*)
4. Моделирование (*Modeling*)
5. Оценка (*Evaluation*)
6. Внедрение (*Deployment*)



Практические навыки

- импорт-экспорт данных,
- проверка и исправление типов данных,
- кодирование категориальных переменных,
- переименование-добавление-удаление 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.

- Вырабатывать критическое отношение к данным.
- Давать возможность студентам испытать сложность работы с большими, комплексными, беспорядочными датасетами.
- Искать паттерны и закономерности, уметь менять масштаб рассмотрения проблемы
- Исследовать, как разные предназначения одних и тех же данных влияют на их представление и трансформацию

Дата Кампус: программа

Анализ данных и машинное обучение в решении социально-экономических проблем
Введение от экспертов, погружение в тему, знакомство с датасетами.

1

2

Python для анализа данных и машинного обучения
Знакомство с языком программирования Python и его специализированными библиотеками для работы с данными и для машинного обучения. Практические упражнения и решение аналитических задач по темам групп с использованием правительственных данных.

3

Проектирование средств управления в социально-экономической сфере на основе данных
Работа с реальными данными, построение гипотез, разработка моделей, консультации с экспертами.

Дата Кампус: предметная программа

Программирование на Python для анализа данных:

- Основы машинного обучения и основные типы задач. Классификация задач машинного обучения.
- Знакомство со специализированными библиотеками Python для научных расчетов и анализа данных.
- Знакомство с различными методами предобработки данных, описательными статистиками и основными способами визуализации данных.
- Методы обучения с учителем и без. Методы кластеризации. Критерии оценки качества полученных моделей. Линейный регрессионный анализ.
- Задача классификации и основные методы ее решения. Метрики качества классификации.
- Введение в Deep Learning.

Всего: 44 часа

Теория: 15 часов

Практика: 29 часов

Проектная работа

- Правоохранительная система;
- Здравоохранение;
- Культура;
- Наука и образование;
- Экология и природные ресурсы;
- Экономика, промышленность и торговля;
- Сельское хозяйство;
- Транспорт и энергетика;
- Финансы;
- Коммуникации, связь и информационные технологии

- Каждая группа должна сформулировать обоснованную гипотезу с использованием некоторых параметров (категорий данных, фич), раскрывающую некоторый аспект социально-экономической реальности. Выявление связи, закономерности, тренда или иное. Ваша ГИПОТЕЗА должна касаться профиля «лаборатории».
- Вы можете использовать любые данные из любых областей, включать в свой датасет любые признаки, находить любые зависимости, строить любые модели.
- Средствами анализа данных вы должны подтвердить или опровергнуть свою гипотезу.

+ Конкурс студенческих проектов «Управление на основе данных»

Учебные задачи

- Социо-экономическая статистика: дата-майнинг, предиктивная аналитика
- Анализ текстов СМИ, «цифровая гумани-таристика», методы обработки естественного языка;
- Распознавание образов, цифровые экосистемы

  (классификация, регрессия, кластеризация).

- Задачи варьируются по тематике, применяемым алгоритмам, степени сложности
- Проблемные, практикоориентированные, не имеющих готового ответа, дробящиеся в своей структуре на подзадачи
- Следуют принципам:
  - релевантности;
  - аутентичности и дозированной сложности данных.

“Национальная стратегия развития искусственного интеллекта на период до 2030 года”

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

Указ Президента Российской Федерации от 10.10.2019 г. № 490 О развитии искусственного интеллекта в Российской Федерации / Сайт Президента РФ. URL: http://www.kremlin.ru/acts/bank/44731

Дата Кампус: проекты

Проект	Краткое содержание	Освоенные методы	Тематика
Антикаракули	Работа посвящена распознаванию рукописных кириллических символов с использованием машинного обучения. «Актуальность: большое количество рукописных документов, имеющее культурное значения для человечества, не оцифровано». Результат: построенная модель распознает рукописный текст с точностью 92%.	Классификация. Нейросеть; keras	Коммуникации; Цифровые экосистемы
Экономическая рокировочка	Предиктивная аналитика. Предсказано повышение стоимости потребительской корзины в РФ в 2020 году (признаки в модели: Цена зерна, Средняя Зарплата, Цена хлеба, Инфляция, Индекс потребительских цен, МРОТ, прожиточный минимум).	Нейросеть, Multi Layer Perceptron Regressor	Экономика
Сортировка мусора	Сортировщик для раздельного сбора мусора. Построена нейросеть, классифицирующая изображения мусора на 6 категорий с точностью 69.2%; “мы научились: Составлять собственный датасет; Создавать сверточную нейронную сеть; Настраивать сеть, загружать веса, изменять модель, приспосабливая её под многоклассовую классификацию; Проводить настройку слоёв оптимизации нейронной сети; Оценивать работу модели.”	Классификация. Нейросеть; keras	Экология
Global Terrorism	Используя датасет с данными о 180,000 атаках, произошедших во всем мире с 1970 по 2017 г, предпринята попытка предсказания цели нападения. Применена визуализация с нанесением данных на карту мира.	RandomForestClassifier, LogisticRegression, matplotlib	Безопасность

Дата Кампус: проекты

Проект	Краткое содержание	Освоенные методы	Тематика
Добывающая промышленность	На основе статистики инцидентов (26 признаков, 350 тыс сэмплов) в горнодобывающей отрасли в США с 1991 по 2018 год построена модель, выделяющая факторы травмоопасности.	Кластеризация, KMeans	Промышленность
Пенсионеры	Вычисление пенсионного возраста. Собрали датасет по 37 странам и 9 признакам (популяция; ВВП на душу; Сред.ЗП; Ср. кол-во детей; Средний возраст; Прод. жизни; % пенс; Размер пенсии; Пенс. Возр.), построили 4 модели, определили отклонение пенсионного возраста в РФ относительно других стран в выборке.	Линейная регрессия, нейросеть, дерево решений, K-neighbors	Экономика; Социология
Энерджи (aka Solar Moon)	На данных действующей солнечной электростанции, получена зависимость мощности СЭС от различных погодных условий; рассмотрены варианты рекомендаций по географическому применению солнечных панелей для генерации электроэнергии.	методы линейной регрессии, случайного леса, ближайших соседей	Энергоносители; Экология
Урбанистика	Анализ времени ожидания общественного транспорта в городах РФ. Организовали интернет-опрос, собрали датасет из 30 тыс сэмплов, 7 признаков. Установили связь между стоимостью проезда и временем ожидания. Интерпретировали связь между удовлетворенностью работой общ.транспорта и досупом частных перевозчиков к оказанию услуг в городе.	LinearRegression, KNeighborsRegressor, MLPRegressor	Муниципальное управление; Транспорт
Пропажа людей	Цель - привлечь внимание общественности к проблеме пропажи людей. На основе статистики МВД собрали и проанализировали данные по регионам РФ о пропаже людей за 5 лет, выделили регионы с высоким показателем. Построили модель	Линейная регрессия	Безопасность; Муниципальное управление; Право

Дата Кампус: результаты

- Предметный результат:
  знакомство с основными методами, техниками и задачами анализа данных и машинного обучения.
- Образовательный:
  пробно-проектная деятельность, связанная с анализом реальных проблем управления в социально-экономической сфере, и выработка сценариев управленческого решения этих проблем.
- Личностный результат:
  формирование представления о профессии Data Science и самоопределение по отношению к карьере в этой области.