5. Машинное обучение [Электронный ресурс] // URL: https://ru.wikipedia.org/wiki/Машинное_обучение (дата обращения 03.04.2021)

Костенко В.В.

Как (не) надо делать учебные программы по прикладному анализу данных. Опыт ПАНДАна

*Европейский университет
(Россия, Санкт-Петербург)*

doi: 10.18411/trnio-11-2021-39

Аннотация

В сентябре 2020 года Европейский университет совместно с «Яндексом» запустил программу дополнительного образования «ПАНДАН» («Прикладной анализ данных»). Цель этой программы – научить людей без математической подготовки применять языки программирования и машинное обучение к реальным задачам из таких областей как история, история искусств, этнография, антропология. Сейчас идет четвертый модуль программы, за это время стало понятно, какие методические и организационные решения были верными, а какие – ошибочными. Перед запуском магистратуры мы хотели бы обсудить эти решения с коллегами, многие из которых также решают задачу обучения своих студентов количественным методам без математической базы.

Ключевые слова: прикладной анализ данных, статистика, проектное обучение, методические рекомендации.

Abstract

European University at St.Petersburg and Yandex, an IT-giant, have launched an educational program ADA/PANDAN (“Applied Data Analysis”) in September, 2020. The aim of the program is to teach people with no math knowledge to apply programming languages to real tasks derived from history, art history, ethnography, and anthropology. As we are finishing the 4th module of the program, we now see which of our methodical and organizational decisions were right or wrong. Before launching a Master program based on the same idea and principles, we want to share and discuss our experience with those colleagues who also teach statistics and data analysis to students without mathematical background.

Keywords: applied data analysis, statistics, project learning, guidelines.

Введение

Профессии аналитика данных, программиста, специалиста по машинному обучению входят в топ-5 наиболее востребованных специальностей на рынке труда, недостаток этих специалистов как с базовыми, так и с более продвинутыми компетенциями растет вместе с цифровизацией мировой экономики. Сейчас аналитики данных нужны не только в IT, но и в корпорациях, банках, на производстве. Если еще несколько лет назад аналитикам было достаточно знать статистику и визуализировать результаты, сейчас требуется опыт работы с большими данными, машинным обучением, нейросетями.

Движение в эту сторону происходит и со стороны государства. В марте 2021 года вице-премьер Дмитрий Чернышенко поставил задачу закрыть дефицит кадров по IT-специальностям в ближайшие годы, для этого нужно подготовить от 700 000 до 1 000 000 специалистов. До 2024 года нужно подготовить около 80 тысяч преподавателей по цифровым специальностям, уже 16 тысяч в текущем году.

Открывается немало бакалаврских и магистерских программ, где обучают этим профессиям. Однако эти программы рассчитаны, в первую очередь, на выпускников школ. В то же время многим взрослым, в том числе состоявшимся специалистам, очевидно, что эти навыки необходимы [1], но вариантов обучения для них мало: это либо онлайн-курсы, либо очные магистерские, что часто невозможно для работающих людей.

В такой ситуации вечерняя программа дополнительного профессионального образования оказывается очень востребованной. В 2020 году конкурс на ПАНДАН составил 6 человек на место при отсутствии большой рекламной кампании. Однако этот спрос и исключительно высокая мотивация студентов – это серьезный вызов для организаторов, особенно в первый год.

В этой статье мы хотим рассказать о том, с какими изначальными представлениями мы создавали эту программу, и как эти представления изменились в ходе работы со студентами.

История вопроса

Попытки научить студентов гуманитарных и социологических специальностей программированию и анализу данных начались еще в 1990-е годы, а сейчас идет уже третья волна интереса к этой идее. Первая волна – это интеграция предмета «Информатика» в школьные программы, изучение языков Basic и Pascal в школах и университетах, а также возникновение Фортрана и последовавших за ним высокоуровневых языков программирования. Программисты получили возможность писать вычисляемые выражения "как есть".

"Вторая волна" началась с двух сторон: появлялись специализированные пакеты программ с элементами программирования, например, SPSS и Excel; дальнейшее развитие получили языки программирования общего назначения, интегрируя в себя парадигмы структурного, функционального и объектно-ориентированного программирования. На этом этапе специалисты из предметных областей и программисты пользовались разными языками программирования и средами. Но все чаще высказывалась мысль про то, что со временем, когда языки программирования станут все удобнее и проще в освоении, программисты разделятся на "системных программистов", которые пишут библиотеки, занимаются оптимизацией и т.д. и на "прикладных программистов", в роли которых все чаще будет выступать человек любой профессии, которому нужна локальная автоматизация в его работе.

После появления динамических интерпретируемых языков (Perl, Python) в сочетании с ростом производительности процессоров и оперативной памяти, стало возможным писать небольшие декларативные программы, используя готовые библиотеки из сотен тысяч доступных. Программист на этом уровне не заботится о выделении и освобождении памяти, типах переменных, сам почти не создает своих функций и классов, почти не использует циклов (они спрятаны внутри библиотек).

В этот момент языки программирования общего назначения практически догнали по удобству специализированные среды. Назовем это "третьей волной", когда написать запрос в библиотеке pandas в Python почти так же просто, как в Excel. Специализированные среды тоже развиваются. Сейчас большой сегмент рынка занимает RPA (Robotic process automation), когда любой человек может создавать программы по локальной автоматизации на своем рабочем месте. Такие решения имеются у многих компаний, включая IBM и Microsoft.

Вместе с тем обучение статистике и анализу данных остается одной из сложнейших тем для студентов, особенно для тех, кто не выбрал эти дисциплины основной специальностью и не имеет сильной математической подготовки. Еще в 1991 году Роберт Хогг [2] отметил, что большинство студентов считают статистику самым скучным и трудным предметом в своих университетских программах и призвал к срочному изменению модели преподавания. Сандра Филлебраун и Томас Лав [3, 4] уже в 1990-е годы утверждали, что даже базовые курсы по статистике следует переводить в проектную работу, а Гэри Смит [5] постулировал необходимость курса по statistical reasoning (что можно перевести как «статистическое рассуждение»). Джонсон и Дасгупта [6] в течение нескольких лет интервьюировали студентов первых курсов по поводу предпочтительного стиля изучения этих дисциплин и показали, что визуальное сопровождение, прикладные занятия и работа в малых группах принципиально увеличивают вовлеченность и удовлетворенность студентов, а также их успеваемость. Дэниэл Шафер и Фред Рэмси [7] пришли к схожим выводам, когда

превратили свой годичный курс по статистике в среду совместной работы над сложными и интересными задачами, в которой студенты испытывали постоянный интерес и потребность осваивать новые инструменты не ради оценки, а для выполнения прикладных задач высокого уровня.

Обращаясь к описанному выше международному опыту, мы как создатели программы стремились создать образовательную среду, в которой инструменты (такие как статистика и языки программирования) будут первым шагом к амбициозной цели, к созданию в небольшой команде настоящего востребованного академического или прикладного продукта, полезного людям. По изначальной задумке, такая цель должна была нивелировать трудности вхождения в новую для студентов область.

Предпосылки

На 2020 год Яндекс уже имел несколько успешных партнерств с вузами, однако это были вузы с факультетами или кафедрами компьютерных наук (например, НИУ ВШЭ, РЭШ, СПбГУ, УРФУ, БГУ), что значительно облегчало интеграцию и создание учебных программ и совместных лабораторий.

В Европейском университете были отдельные центры и факультеты, где преподавались и использовались в научной работе анализ данных и статистика. При этом Европейский университет – это преимущественно гуманитарный вуз с семью факультетами (антропологии, истории, истории искусств, социологии, политологии и экономики + центр философии Стасис). Было изначально понятно, что проникновение цифровых подходов и аналитики данных на эти классические факультеты будет сопряжено с трудностями.

Было принято решение использовать проектный подход и идти от потребностей ученых внутри вуза, чтобы студенты новой программы могли решить для них какие-то прикладные аналитические задачи, а коллеги увидели преимущества качественной аналитики и визуализации.

Отбор абитуриентов

Важным решением, которое пришлось принимать еще до набора студентов, был вопрос вступительных испытаний, а также нагрузки и времени занятий. Мы решили позиционировать программу как возможность для взрослых людей с нетехническим образованием за год освоить новые навыки. Соответственно, мы ожидали работающих людей, а значит занятия можно было ставить только вечером и не задавать много на дом.

Для отбора абитуриентов мы выбрали творческий, но трудозатратный для экзаменаторов путь. Каждый четверг в течение шести недель, начиная с апреля, мы публиковали в социальных сетях Европейского университета одно задание с хэштегом #Панданназавтрак (для допуска к собеседованию нужно было выполнить два из шести заданий). Каждое задание было либо связано с одним из будущих проектов, либо с одним из курсов, и не предполагало одного ответа. Это был вопрос (или история, или фотографии, или графики), на который нужно было дать комментарий в виде небольшого эссе. Никаких количественных или статистических компетенций для этого не требовалось, только умение искать информацию по малознакомой теме и грамотное изложение мыслей.

Всего пришло около 100 анкет, в которых абитуриенты выражали интерес, около 70 прислали оба эссе, немногим меньше 50 были приглашены на собеседование. Собеседование проходило в свободной форме, каждое занимало не меньше получаса. Мы считаем, что такой способ отбора был удачным, поскольку из 26 отобранных абитуриентов на программе до сих пор учатся 23, при этом двое отчислились по личным обстоятельствам и только один не справился с программой.

Отбор преподавателей и кураторов проектов

Перед запуском первого набора мы как организаторы встретились с представителями всех факультетов и одиннадцати научных центров. Некоторым исследователям наше предложение о сотрудничестве не было интересно в принципе, другие не видели перспектив, поскольку не думали в терминах «данных» (например, банк фотографий многие исследователи не воспринимают как данные). Однако несколько центров и рабочих групп с

энтузиазмом откликнулись, и нам быстро удалось придумать прототипы четырех проектов. Так были найдены кураторы со стороны Европейского университета, которые отвечают за содержательную повестку.

Со стороны Яндекса техническое кураторство двух проектов осуществляет со-руководитель ПАНДАНа, двух других – приглашенные специалисты, интересы которых близки к тем техническим компетенциям, которые развиваются в проекте (оба они – сотрудники других университетов, один также преподает курс на ПАНДАНе). Никакого специального тренинга для кураторов мы не проводили, что было ошибкой.

Для чтения курсов мы искали ведущих специалистов в каждой области, сумев собрать сильную команду. Такой подход имел как очевидные достоинства, так и серьезные недостатки, о чем подробнее см. ниже в разделе «Что пошло не так».

Перед началом набора мы провели ретрит, на котором собрали преподавателей программы и кураторов курсов. Эта рабочая группа помогла лучше понять подход каждого преподавателя, отказаться от некоторых курсов, которые плохо укладывались в общий план, выстроить курсы по модулям, познакомить кураторов. Кроме того, мы старались донести до преподавателей, особенно технических дисциплин, что уровень изначальной подготовки студентов будет ниже с точки зрения математики, чем тот, с которым они привыкли работать.

Реализация

Команда

В ядре команды с самого начала был один со-руководитель от Яндекса, один от Европейского университета, а также исследователь, применяющий количественные методы в своей научной работе, и администратор, заинтересованный в том, чтобы также стать студентом программы. Такая команда позволила быстро выходить на связь с университетскими преподавателями и сотрудниками Яндекса, плотно координировать проектную работу, интегрировать проекты между собой, а также понимать, как чувствуют себя студенты на том или ином курсе в режиме реального времени, а не только на момент проведения опросов. Это взаимодействие также давало возможность проговаривать ожидания всех сторон, обсуждать реалистичность планов, помогать друг другу, когда ситуация разворачивалась по неожиданному сценарию.

Курсы

Изначально планировалось провести 14 курсов с тем, чтобы студенты могли выбрать из них 11. В ходе учебы преподаватели одного из курсов поняли, что их изначальная концепция плохо связана с основным направлением программы и свой курс сняли. В итоге на программе осталось 13 курсов, из которых выбрать нужно было 10. На ПАНДАНе принята модульная система, где 1 модуль равен 7 неделям, в каждом семестре (один из которых называется «Вызовы», другой – «Ответы») 2 модуля. Между семестрами каникулы составляют месяц, между модулями – неделю. Половину курсов читали преподаватели Европейского университета, другую – специалисты из Яндекса.

В 2010 году Нолан и Ланг [8] призвали полностью перевести все занятия по статистике и анализу данных в среды программирования или статистические пакеты, а также включать модули или целые курсы по анализу данных в университетские программы, где раньше преподавалась только общая статистика. Мы были полностью согласны с этой идеей, которая за последние 10 лет превратилась из радикальной в очевидную. В первом модуле все курсы были обязательными и формировали общую базу: технический минимум, язык программирования Python, программирование в R и введение в статистику и вызовы в общественных науках. Во втором модуле обязательным был только курс по машинному обучению, курсы по типам данных (административные, опросные и текстовые) были выборными.

В третьем модуле одновременно шли 3 курса по выбору: разметка данных, компьютерное зрение и обработка естественного языка, а также началась работа над проектами (им посвящался один вечер). В четвертом модуле студенты должны взять

обязательный курс по визуализации данных [9] и могут выбрать курс по причинно-следственным связям в неэкспериментальных данных.

Проекты

Работа с реальными данными в рамках четырех проектов изначально была самой привлекательной частью курса для большинства студентов, которая удерживала их уровень мотивации во время самого напряженного первого модуля [10]. У каждого проекта был содержательный куратор от Европейского университета и технический от Яндекса, а также дополнительная договоренность с Яндексом о возможности точечного подключения узких специалистов для консультаций. Некоторые проекты были запущены в работу команду (при поддержке стажеров из Computer Science Center) заранее, чтобы не допустить прихода студентов в неподготовленный проект, в котором трудности технического и содержательного характера не позволят в итоге его успешно реализовать.

Особенно это было важно для проекта «Госкаталог», в котором в отличие от трех других не было человека, для которого его научные амбиции завязаны на конечном продукте. Этот проект, изначально запущенный министерством культуры для оцифровки всего многомиллионного музейного фонда России (<https://goskatalog.ru/portal/>), вызывает большое недовольство у сотрудников музеев, поскольку значительная часть их рабочего времени в последние годы связана с авральным внесением данных обоих фондов в эту систему. При этом неудачные решения при вводе (нет никаких фильтров и проверок) приводят к тому, что пользоваться поиском по той четверти фондов, которые уже оцифрованы, просто невозможно. Сложившееся за годы плохое отношение профессиональных искусствоведов к Госкаталогу не позволило нам найти на факультете истории искусств ЕУСПб интересанта проекта, но одна из коллег, которой много приходилось с ним работать, согласилась консультировать команду.

В остальных трех проектах, напротив, куратор со стороны Европейского был держателем собственных данных, а также видел научную цель в том проекте, который реализовывала команда.

Так в проекте «Бомбардировки Ленинграда» содержательным куратором стал историк, который несколько лет собирал в архивах данные МПВО (Местной противовоздушной обороны), в которых скрупулезно зафиксирован каждый налет и бомбардировка Ленинграда с количеством жертв и описанием разрушений. Эти уникальные данные при удачной визуализации способны рассказать историю блокады день за днем. А привязывание по геометкам тех 200 блокадных дневников, которые коллекционирует проект «Прожито» к карте бомбардировок позволит создать еще более полное полотно жизни того времени.

Коллаборация с проектом «Прожито», однако, заключалась не в блокадных дневниках (эта идея возникла уже в ходе реализации). Здесь идеей было создание на основе нейросети GPT-3 инструмента, воспроизводящего язык и эмоции подростков. Прототипом был бот в телеграме, который позволил бы искать фразы по некоторым темам (например, «несчастливая любовь», «плохая оценка», «отношения с родителями»), но в ходе реализации возникла идея обучить нейросеть на типичных репликах подростков разных эпох, чтобы современные школьники могли полноценно «поговорить» со своими предшественниками.

Еще одним проектом, который оказался очень популярным, стали «Еврейские надгробия». Центр «Петербургская иудаика» в Европейском университете хранил уникальную коллекцию надгробий, сфотографированную в 1930-е годы Давидом Гоберманом. Их автоматическая разметка и расшифровка на базе инструмента, обученного на коллекции расшифрованных надгробий, предоставленных центром научных работников и преподавателей иудаики «Сэфер», и стала задачей проекта.

Каждый проект был связан с некоторыми курсами и методами работы с данными. «Госкаталог» основан на курсах «Машинное обучение» и «Разметка данных» (а курса по базам данных для него явно не хватило), «Бомбардировки Ленинграда» - это, в первую очередь визуализация (и здесь не хватило курса по геоданным). «Прожито» основывался на

курсах «Текстовые данные» и «Обработка естественного языка», «Еврейские надгробия» - на курсе «Разметка данных», «Машинное обучение» и «Компьютерное зрение».

Что пошло не так.

— **Линейная алгебра и математический анализ.**

Поскольку идея проекта изначально была в том, что абитуриенты приходят без математической подготовки, их уровень математики почти во всех случаях был хорошо забытым школьным. Поскольку мы не предусмотрели курс по линейной алгебре или математическому анализу, вспоминать даже довольно простые вещи из школьного курса было тяжело, и часто на это тратилась немалая часть занятий на таких курсах как, например, машинное обучение. Мы думали обойтись только прикладными дисциплинами, но довольно быстро стало понятно, что такое решение просто перекладывает задачу вспоминания математики на преподавателей профильных предметов.

— **Базы данных.**

Даже современные выпускники технических специальностей нередко недостаточно знакомы с устройством и поддержанием сложных баз данных, отдавая предпочтение более модным направлениям, таким как машинное обучение и нейросети. При этом реальная потребность в таких навыках очень высока как в реальном секторе, так и в тех проектах, которые реализуются на ПАНДАНе. Отсутствие этого курса мешало, поэтому пришлось его заменять лекциями ad hoc.

— **Очень высокая нагрузка в 1 модуле, особенно домашние задания.**

4 вечера в неделю занятия шли с 18.30 до 21.00, а до этого большинство людей работали. Хотя в изначальном дизайне так и было задумано, преподаватели, намного более привычные к студентам, учащимся в дневное время, не сразу смогли перестроиться. Поскольку у преподавателей было всего 7 сдвоенных пар на курс (28 академических часов), многие старались перенести часть образовательного процесса на самостоятельную работу и давали домашнее задание на каждую неделю, причем по некоторым курсам оно занимало от 10 до 24 часов. Это быстро привело к первому кризису, в ходе которого мы построили систему обратной связи через опросы, а в отдельных случаях и с помощью личных бесед. При этом некоторым преподавателям пришлось перестраивать курсы на ходу, что всегда сложно.

— **Лучшие преподаватели – это очень занятые люди.**

Как уже отмечалось выше, решение пригласить очень высокоуровневых преподавателей имело как свои положительные, так и отрицательные последствия. Поскольку профессионалы в IT всегда очень заняты, нередко они не могут посвящать много времени преподаванию. Студенты ПАНДАНа же требовали очень значительных временных ресурсов преподавателей из-за того, что их базовой подготовки часто не хватало для того, чтобы успевать усваивать материал. В результате иногда оказывалось, что материал, который рассказывал младший научный сотрудник или аспирант, воспринимался легче, чем лекции от именитых лекторов, которым трудно было представить, как сложно начинать освоение анализа данных с нуля. Мы пришли к выводу, что специалистам высокого уровня трудно понять, как студенты могут не знать базовых вещей, не говорить с ними на одном языке, так как опыт преподавания у них в основном для студентов с блестящей математической подготовкой.

— **Отсутствие достаточной инфраструктуры, а также базового центра \ лаборатории для поддержки проектов.**

Когда мы планировали программу, мы не могли предвидеть, какими масштабными станут проекты, над которыми мы работаем. Чтобы не тормозить процесс обучения из-за разных машин слушателей, мы выделили каждому виртуальную машину [11]. В результате работа с серверами легла на плечи одного из организаторов, потому что мы не предусмотрели своевременный найм системного администратора. Большая нагрузка на организаторов была также связана с тем, что у нас отсутствовала лаборатория (или центр),

сотрудники которой занимались бы текущей поддержкой работы команд как с технической, так и с менеджерской точки зрения. На данном этапе мы паллиативно решили эту проблему назначением менеджеров проектов из среды студентов и обучением их базовым принципам agile и использованию Git для контроля версий [12]. В новом учебном году поддерживать проекты будет отдельный научно-прикладной центр МАСТ, в котором будут трудоустроены программисты и дизайнеры. Мы надеемся, что наличие такой инфраструктурной поддержки снимет сверхнагрузку с руководителей программы и кураторов.

— **Работа с людьми из бизнеса, а не из академии.**

Когда мы запускали образовательную программу, мы были уверены, что как минимум половина абитуриентов будет связана с бизнесом. Эта гипотеза не подтвердилась, большинство студентов тем или иным образом связаны с гуманитарным знанием и академической средой; с ними мы говорим на одном языке. Однако для людей, уже реализовавших себя в бизнесе, отсутствие небольших конкретных задач, долгие циклы, много учебы, стресс от того, что не получается сразу, оказались очень некомфортными. Мы пришли к выводу, что пока не умеем работать с людьми из бизнеса.

Что удалось

— **Набрать очень мотивированных студентов.**

Оказалось, что обучение анализу данных на примерах и проектах из гуманитарных наук очень востребовано. Таких программ не хватает, поэтому мы могли себе позволить строгий отбор. Кроме того, не будучи уверенными в успехе интеграции Европейского и Яндекса, мы поставили низкую цену (100 тысяч рублей за весь восьмимесячный курс), это привлекло еще больше абитуриентов. В результате студенты очень хотели учиться, были готовы мириться с неудачами своими и организаторов, открыто строили диалог о своих проблемах, подстраивали свои графики под учебный план. Учитывая все трудности, с которыми они столкнулись, особенно в 1 модуле, когда количество заданий было несоизмеримо с тем временем, которым они располагали, отчисление трех человек из 26 мы считаем отличным результатом.

— **Высокое качество курсов.**

Как уже отмечалось выше, для преподавания курсов были ангажированы преподаватели высокого уровня. Некоторые из них смогли очень быстро подстроиться под уровень студентов и обеспечить сверхбыстрый процесс обучения. Другие коллеги именно в силу своего высокого уровня потратили немало времени, чтобы приспособиться к базовому уровню студентов. Двое преподавателей были сильными аспирантами, очень заинтересованными в своем развитии как преподавателей и в деле программы. Этот удачный опыт мы планируем развить через систему учебных ассистентов, чтобы обучить новых преподавателей под нужды непосредственно нашей программы, поскольку сейчас этот путь по многим дисциплинам выглядит проще, чем интеграция существующих педагогов в те условия, которые мы предлагаем.

— **Наладить эффективную обратную связь.**

В конце каждого модуля (а в первом модуле в середине и в конце) студентам предлагалось заполнить опросник, где задавались как закрытые вопросы о времени, которое они тратят на домашние задания, пользе от каждого курса по ряду критериев, так и открытые о том, с какими трудностями они сталкиваются, что нужно поменять. После сбора опросников мы каждый раз обсуждали результаты опроса на очной встрече, открыто признавали свои недоработки и стремились улучшить те аспекты, о которых говорили студенты. Наличие этого постоянного диалога, хотя и требовало значительной вовлеченности, решало многие проблемы и создавало коллективный дух и общую ответственность. Дополнительными инструментами, позволяющими отслеживать текущую ситуацию стали трекер и несколько чатов в телеграме (по курсам, по проектам и по проблемам).

— **Наладить интеграцию между «физиками» и «лириками», придумать и запустить проекты, которые интересны всем сторонам.**

Самой амбициозной задачей ПАНДАНа было включение студентов с полугодовой подготовкой в реальные академические проекты. Причем эти проекты тоже необходимо было запустить, потому что три из них не включали в себя it-компонент, а четвертый (улучшение интерфейса Госкаталога) просто не существовал до того, как мы его придумали. Сейчас на этапе начала четвертого модуля можно уверенно сказать, что команды сложились и работают. Исследователи сначала относились к аналитикам с некоторой опаской, но со временем увидели, какие возможности открывает это сотрудничество, и стали быстро увеличивать свои запросы, а вместе с их реализацией стал расти и энтузиазм. Сейчас мы столкнулись скорее с проблемой нехватки ресурсов (человеческих и инфраструктурных) для реализации всех исследовательских амбиций. Эту проблему мы планируем решать с помощью создания центра при программе и с помощью привлечения практикантов из Computer Science центра – другой образовательной программы Яндекса с JetBrains.

— **Интегрировать студентов Computer Science Center в проекты.**

Одной из возможностей для студентов ПАНДАНа, но одновременно и одной из задач для организаторов была возможность привлекать практикантов из Computer Science центра на 2-3 месяца под определенные проекты для помощи в тех областях, где ПАНДАН компетенций не дает (например, создание баз данных, фронтэнд или бэкэнд). При этом мы также планировали создать для программистов место, где есть интересные, не оторванные от жизни задачи, и где все студенты могут учиться взаимодействию в команде, приближенному к реальной жизни. Этот опыт по большей части был очень успешным.

О чем еще нужно думать.

Некоторые аспекты работы программы остаются для нас спорными или нереализованными, именно здесь мы нуждаемся в дискуссии с профессионалами в области высшего образования и преподавания анализа данных.

— **2 языка программирования.**

В данный момент обучение идет на языках R и Python, поскольку первый из них используется сотрудниками Европейского университета и в академической аналитике данных, а второй – в коммерческом программировании. Обучение двум языкам тяжелее для студентов, но проще для преподавателей. На первом году было принято решение в пользу упрощения подготовки курсов для преподавателей, поскольку им и так пришлось радикально упрощать свои курсы для того, чтобы сделать их доступными для студентов без изначальной технической подготовки. Вместе с тем, в долгосрочной перспективе такое решение выглядит неоптимальным. Для того, чтобы выйти из этой ситуации, нужно переобучать преподавателей новому языку.

— **Боты для обучения азам языков.**

Как предупреждают Донохью и коллеги [1], восприятие языков программирования и базовой статистики исключительно как ступеньки к более интересному анализу данных часто ведет к тому, что эти курсы читаются довольно скучно, им не уделяется должного внимания, а преподаватели оставляют значительную часть материала на самостоятельное прохождение. Для людей, столкнувшихся с этими предметами впервые, это опасная тенденция, часто ведущая к грубым ошибкам впоследствии. На ПАНДАНе сжатый цикл обучения не дает возможности долго заниматься со студентами языками программирования, однако без этой основы невозможно дальнейшее движение вперед. Поэтому мы запустили экспериментальный проект – телеграм-бот для обучения языку, где в игровой форме студент еще до начала занятий сможет ознакомиться с базовыми функциями Python. Мы планируем, что интерес к работе с этим ботом появится у студентов в ходе прохождения ознакомительного онлайн-курса по Python, который станет пререквизитом для поступления на программу в следующем году.

— **Встраивание студентов и кураторов в проекты.**

На первом году мы предлагали людям выбрать один из четырех проектов, знакомили их с кураторами, а затем подключались только в случаях, когда была нужна помощь. Сейчас

мы видим, что этого было недостаточно. Необходима более плотная работа с кураторами, возможно, их специальное обучение, а также более внимательная работа с каждым студентом на этапе выбора проекта, чтобы она или он лучше понимали цели проекта, методы, которые в каждом из них будут изучены, а также промежуточные задачи.

— **Метод управления проектами.**

В современном IT Scrum стал одним из ключевых методов управления проектами, однако мы использовали только его элементы. Возможно, введение полноценного управления через scrum снимет часть трудностей не только с администраторов, но и с кураторов, позволив лучше понимать процессы и направлять их.

— **Чрезмерная нагрузка и проблема перегорания.**

Как и многие стартапы, наш проект в первый год во многом был успешен за счет постоянных переработок всех членов команды, их готовности идти навстречу, откладывая другие дела для решения проблем на проектах и курсах ПАНДАНа. Такая модель нежизнеспособна в долгосрочной перспективе, хотя в краткосрочной способствует большей сплоченности и вовлеченности, эйфории от успехов, а у студентов появляется ощущение очень персонализированного подхода и заботы о них. Для того, чтобы не допустить перегорания всех вовлеченных людей нам кажется необходимым четче разделить обязанности, нанять несколько дополнительных сотрудников и выстроить инфраструктуру под потребности программы. На начальном этапе это, как правило, невозможно, но в случае успеха проекта просто необходимо, поскольку иначе проект не сможет существовать долго.

— **Критерии успешности.**

Сейчас мы видим три основных критерия успешности программы:

1) Востребованность выпускников на рынке труда; 2) Заинтересованность экспертов и исследователей в получившемся продукте; 3) Популярность программы у абитуриентов.

Это совсем не исчерпывающий список, а скорее то, на что мы фактически опирались в работе. Вполне возможно, что это достаточные цели на начальном этапе, но в будущем необходимо более детальное и стратегическое целеполагание.

— **Как масштабировать проект?**

В текущем виде проект ПАНДАН – это история совмещения гуманитарных дисциплин и IT, скорее digital humanities. При всей интересности этой сферы, круг заинтересованных в ней людей не так уж широк. Для того, чтобы сделать программу по-настоящему привлекательной и масштабируемой, мы полагаем, что нужно увеличить число проектов, связанных с более прикладными, а не академическими задачами. Например, в следующем году мы запускаем новый проект «Безопасные дороги», который объединит данные ГИБДД и госзакупок по нацпроекту, из которых видно, какие дороги ремонтировались в последние годы. Это позволит показать эффективность использования средств и выяснить, действительно ли отремонтированные дороги стали безопаснее с точки зрения аварийности и смертности. Такие проекты интересны не только академическим исследователям, но и бизнесу, и некоторым государственным структурам, использующим анализ данных. Нам кажется, что такого рода проектная деятельность уже может быть лучше масштабируемой, пускай не на сотни студентов бакалавриатов крупных вузов, но хотя бы на десятки магистрантов.

Заключение

Еще с 1990 – х годов проблема скучного и перегруженного математикой преподавания статистики и анализа данных обсуждается профессионалами [2]. На программе ПАНДАН мы попытались использовать те наработки, которые описаны в современной методической литературе, чтобы создать курс по анализу данных для максимально широкой аудитории, сделать его доступным и интересным (аналогичную попытку описывают Четинкая и Эллисон [13] на примере своего курса в университете Дьюка). Мы пришли к выводу, что несмотря на все сложности, создание таких курсов возможно и необходимо,

особенно учитывая стремительный рост потребности рынка труда в специалистах - аналитиках данных.

Те трудности, с которыми мы столкнулись, связаны как с инерцией дисциплин, в которых междисциплинарные проекты все еще редки, так и с тем, что многим людям, которые привыкли думать про себя как про «гуманитариев» непросто поверить в свои силы в аналитике [14].

Наши эмпирические результаты подтверждают находки Иванюшиной и коллег [15], показывающих, что ожидание успеха становится одним из ключевых факторов вовлеченности, а его отсутствие и ощущение постоянного напряжения и неудачи связаны с резким падением мотивации и успеваемости. Создание благоприятной инклюзивной среды, где есть возможность поделиться своими впечатлениями, страхами, неуверенностью и успехами, стремление идти навстречу потребностям студентов со стороны преподавателей, становятся ключевыми факторами вовлеченности [16]. Проектная работа в этом случае воспринимается как один из более современных и интересных подходов в работе со сложными темами [17].

В заключение важно сказать про нахождение или формирование общего языка между специалистами из IT-сферы и гуманитариями. В нашем случае трудности перевода были не только и не столько терминологические, но скорее связанные с разными задачами. До тех пор, пока команда не понимает, к какой цели они все вместе двигаются и зачем в команде каждый человек, диалог выстраивается плохо, несмотря на все старания. Однако плотная командная работа с качественным кураторством и обратной связью помогает наладить сотрудничество. Кажется, проговаривание зон ответственности, артикуляция экспертизы и вклада каждого члена команды формирует взаимное уважение и интерес к тому неизвестному, что приносит другой. Такая среда дает возможность не бояться спросить что-то непонятное, выучить новые слова, а иногда и целые области, не чувствуя себя глупым или неподготовленным. В ситуации быстрой демократизации анализа данных именно междисциплинарные команды могут стать той точкой, где знания перетекают наиболее эффективно и создают стимул для дальнейшей учебы.

1. Donoghue, T., Voytek, B., & Ellis, S. E. (2020). Teaching Creative and Practical Data Science at Scale. *Journal of Statistics Education*, 1-22. doi.org/10.1080/10691898.2020.1860725
2. Hogg, R. V. (1991). Statistical Education: Improvements are Badly Needed. *The American Statistician*, 45(4), 342–343. https://doi.org/10.1080/00031305.1991.10475832
3. Fillebrown, S. (1994). Using Projects in an Elementary Statistics Course for Non-Science Majors. *Journal of Statistics Education*, 2(2), null. https://doi.org/10.1080/10691898.1994.11910470
4. Love, T. E. (1998). A Project-Driven Second Course. *Journal of Statistics Education*, 6(1). https://doi.org/10.1080/10691898.1998.11910605
5. Smith, G. (1998). Learning Statistics by Doing Statistics. *Journal of Statistics Education*, 6(3). https://doi.org/10.1080/10691898.1998.11910623
6. Johnson, H. D., & Dasgupta, N. (2005). Traditional versus Non-traditional Teaching: Perspectives of Students in Introductory Statistics Classes. *Journal of Statistics Education*, 13(2). https://doi.org/10.1080/10691898.2005.11910558
7. Schafer, D. W., & Ramsey, F. L. (2003). Teaching the Craft of Data Analysis. *Journal of Statistics Education*, 11(1). https://doi.org/10.1080/10691898.2003.11910692
8. Nolan, D., & Lang, D. T. (2010). Computing in the Statistics Curricula. *The American Statistician*, 64(2), 97–107. https://doi.org/10.1198/tast.2010.09132
9. Loy, A. (2021). Bringing Visual Inference to the Classroom. *Journal of Statistics and Data Science Education*, 0(ja), 1–21. https://doi.org/10.1080/26939169.2021.1920866
10. Libman, Z. (2010). Integrating Real-Life Data Analysis in Teaching Descriptive Statistics: A Constructivist Approach. *Journal of Statistics Education*, 18(1), null. https://doi.org/10.1080/10691898.2010.11889477
11. Kim, B., & Henke, G. (2021). Easy-to-Use Cloud Computing for Teaching Data Science. *Journal of Statistics and Data Science Education*, 29(sup1), S103–S111. https://doi.org/10.1080/10691898.2020.1860726
12. Beckman, M. D., Çetinkaya-Rundel, M., Horton, N. J., Rundel, C. W., Sullivan, A. J., & Tackett, M. (2021). Implementing Version Control With Git and GitHub as a Learning Objective in Statistics and Data Science Courses.

- Journal of Statistics and Data Science Education, 29(sup1), S132–S144. <https://doi.org/10.1080/10691898.2020.1848485>
13. Çetinkaya-Rundel, M., & Ellison, V. (2021). A Fresh Look at Introductory Data Science. *Journal of Statistics and Data Science Education*, 29(sup1), S16–S26. <https://doi.org/10.1080/10691898.2020.1804497>
 14. Fouché, C., & Bartley, A. (2016). Teaching Data Analysis to the Data-Averse: A Framework for Educators. *Journal of Teaching in Social Work*, 36(1), 70–83. <https://doi.org/10.1080/08841233.2016.1125231>
 15. Иванюшина, В., Александров, Д., & Мусабилов, И. (2016). Структура академической мотивации: Ожидания и субъективные ценности освоения университетского курса. *Voprosy Obrazovaniya*, 4, 229–250. <https://doi.org/10.17323/1814-9545-2016-4-229-250>
 16. Wagaman, A. (2016). Meeting Student Needs for Multivariate Data Analysis: A Case Study in Teaching an Undergraduate Multivariate Data Analysis Course. *The American Statistician*, 70(4), 405–412. <https://doi.org/10.1080/00031305.2016.1201005>
 17. Horton, N. J., & Hardin, J. S. (2021). Integrating Computing in the Statistics and Data Science Curriculum: Creative Structures, Novel Skills and Habits, and Ways to Teach Computational Thinking. *Journal of Statistics and Data Science Education*, 29(sup1), S1–S3. <https://doi.org/10.1080/10691898.2020.1870416>

Лавров В.В., Лучкин Р.С., Немыкин О.И., Прохоров М.Е.

Робастный алгоритм оценки параметров прецессии по совокупности оптических изображений

*ПАО Межгосударственная акционерная корпорация «Вымпел»
(Россия, Москва)*

doi: 10.18411/trnio-11-2021-40

Аннотация

Проводится синтез робастного алгоритма определения параметров движения КА в центральном ньютоновском поле сил по совокупности его удалённых снимков, структура которых после обработки одиночных снимков представлена в виде графа леса (совокупности деревьев), восстановленного на этапе обработки последовательности одиночных снимков. Полученные оценки используются для преобразования совокупности снимков в единую объектовую систему координат. Априори предполагается, что распределение сформированной по отдельным снимкам исходной статистики неизвестно и одинаково распределено. Алгоритм основан на поиске М-оценок параметров движения КА как движения твёрдого тела с учетом возможного наличия в исходной статистике «аномальных» замеров. Оценка параметров движения формируется путём последовательного применения алгоритмов, синтезированных с помощью метода наименьших квадратов и поиска медианной оценки на этапе выделения «аномальных» замеров. Приводятся результаты обработки совокупности снимков, полученных при наблюдении КА «Океан-1», при решении задачи оценки параметров регулярной прецессии.

Ключевые слова: М-оценка параметров, медианная оценка, метод наименьших квадратов, модель движения КА, оптическое изображение, регулярная прецессия.

Abstract

A robust algorithm is synthesized for determining the parameters of spacecraft motion in the central Newtonian field of forces from a set of its remote images, the structure of which, after processing single images, is presented in the form of a forest graph (a set of trees) restored at the stage of processing a sequence of single images. The obtained estimates are used to transform the set of images into a single object coordinate system. A priori, it is assumed that the distribution of the initial statistics generated by individual images is unknown and equally distributed. The algorithm is based on the search for M-estimates of the parameters of the spacecraft motion as the motion of a rigid body, taking into account the possible presence of «anomalous» measurements in the initial statistics. The estimation of the motion parameters is formed by the sequential application of algorithms synthesized using the least squares method and the search for the median estimate at the stage of identifying «anomalous» measurements. The results of processing a set of images