ЧТО ТАКОЕ НАУКИ О ДАННЫХ

и почему именно сейчас студентам и аспирантам надо начинать в ней работать

Будучи членом гугл-группы ml-world, я практически каждый день получаю со всех концов мира, от Шанхая до Принстона и Гарварда, объявления о поиске аспирантов, постдоков, профессоров, исследователей по тем или иным аспектам самой востребованной на сегодня области прикладной математики – Data science или Науки о данных (хотя этот перевод, на мой взгляд, звучит немного топорно и что-то неуловимое теряет при всей внешней похожести).

Прежде всего о терминологии. Не надо смешивать Data science и Big data. Это области отчасти коррелированные, но все же очень разные. Если основная цель Big data – это разработка эффективных методов обработки сверхбольших массивов информации, обычно физически отдаленных друг от друга, то цель Data science – извлечение новых знаний из информации, описывающей физические, технологические, социальные и другие процессы. В чем общность этих двух дисциплин? В том, прежде всего, что обе они начинают быть содержательными только в случае, если мы имеем дело с действительно большим объемом данных.

Приведу несколько примеров из очень разных областей человеческой деятельности. Представьте себе, что мы наблюдаем отскоки бесконечно упругого шара от стенок биллиардного стола. Предполагается, что мы не обладаем никакими сведениями из механики – а наша цель состоит в выводе закона, по которому движется шар, на основе сколь угодно длительного наблюдения за координатами двигающегося тела. Входными данными для нас, таким образом, являются декартовы координаты стенок и бесконечный (или лучше сказать сколь угодно длинный) ряд декартовых же координат точек соударения шара со стенками, в которых направление его движения меняется. Можно ли это сделать, то есть найти соответствующий закон, а если можно – то как? Именно не догадаться, как может двигаться шар, а написать программу, которая бы такую закономерность находила без участия человека.

Другой пример: предсказание разного рода катастроф на основе анализа информации из социальных сетей – например, землетрясений или наводнений по изменению поведения домашних животных в определенной местности. С помощью n-грамм (т.е. комбинаций из n слов, описывающих в той или иной мере предсказываемое явление) мы в состоянии свести эту проблему к хорошо известной в математике n-мерной задаче о разладке. Точное теоретическое решение этой задачи неизвестно, но есть довольно много эффективных алгоритмов, решающих эту задачу в некоторых важных частных случаях.

Еще одна задача, на этот раз из банковской сферы. Если нам известна вся статистика поведения клиентов банка, все их транзакции, данные о самих клиентах, известно кто открывал и кто закрывал счет, можно ли с достаточно высокой вероятностью спрогнозировать отток клиентов, идентифицировать тех клиентов, которые могут закрыть счет в недалеком будущем? Оказывается, что такая задача методами Data science вполне удовлетворительно решается.

В общем виде Data science – это наука об извлечении содержательной информации из больших объемов экспериментальных данных, позволяющей определять тренды, находить неизвестные закономерности, а также на основе сочетания методов интеллектуального анализа данных и методов оптимизации находить наборы входных параметров процесса, в наилучшей степени отвечающих заданному функционалу качества этого процесса (так называемая «суррогатная оптимизация»).

Эта наука зародилась всего десять лет назад, прежде всего для применений в области авиации и космоса, как способ определения оптимальных входных данных для процессов, внутренние механизмы которых нам неизвестны (или известны, но слишком сложно вычисляемы), а экспериментальные данные об их функционировании доступны в достаточно большом объеме.
В последние годы стало понятно, что разработанные для этих задач методы применимы в огромном количестве практических задач: в финансовом секторе, страховании, ритэйле, машиностроении, биоинформатике и многих, многих других областях.

Сегодня уже многие десятки университетов за рубежом (и ни одного в России!) выпускают специалистов в этой области. Но нужда в этой компетенции огромна и вы, уважаемые коллеги, не опоздаете, если начнете изучать Науки о данных прямо сейчас.

Заведующий кафедрой ТМСС НИУ ВШЭ
Директор Института проблем передачи информации имени А.А. Харкевича РАН
академик РАН Александр Петрович Кулешов