1 Введение: что такое наука о данных

В последние несколько лет вокруг «науки о данных» или «больших данных» было много шумихи. Первая, вполне обоснованная, реакция на все это — смесь скептицизма и смущения. Действительно, наши реакции (Кэти и Рэйчел) были именно такими.

И мы потакали своим заблуждениям. Сначала каждая сама по себе. А позже вместе, собираясь по средам за завтраком. Однако нас не покидало ощущение, будто в этом *действительно* есть нечто, возможно, глубокое и мудрое, представляющее новую парадигму мышления относительно сферы данных. Вероятно, появилось ощущение, что изменение парадигмы сделает нас сильнее. И вместо того, чтобы игнорировать все это, мы решили разобраться более тщательно.

Но прежде, чем двигаться дальше, разберемся, что именно вызывает сомнения и недопонимание. Вероятно, у вас они тоже есть. Затем мы расскажем, каким образом пришли к заключениям, в результате которых Рэйчел составила курс по Data Science в Колумбийском университете, Кэти опубликовала его в своем блоге, а вы сейчас читаете эту книгу.

Большие данные и наука о данных

Сразу определим наши позиции, так как многие из вас уже испытывают определенный скептицизм в отношении Data Science по тем же причинам, которые были и у нас. Мы хотим сообщить: *мы по одну сторону баррикад*. Если вы тоже скептик, то, вероятно, сможете привнести нечто полезное в процесс легитимизации науки о данных.

Итак, что же вызывает удивление при упоминании больших данных или Data Science? Перечислим главное.

1. Существуют сложности в определении терминологии. Что есть большие данные? Что такое наука о данных? Что общего между первым и вторым? Является ли

Data Science наукой о больших данных? Является ли она всего лишь логическим развитием Google или Facebook, развитием их технологической платформы? Почему многие люди относят большие данные к стыку наук (астрономия, финансы, технология и т. д.), а Data Science считают просто технологией? Насколько велико *большое*? Или это просто удобный термин? Приведенные термины настолько неоднозначны, что почти бессмысленны.

2. Со стороны академической и прикладной наук наблюдается откровенное пренебрежение к исследованиям в этой области, несмотря на то что подобные исследования базируются на десятилетиях (в некоторых случаях — столетиях) работы статистиков, специалистов в области компьютерных наук, математиков, инженеров и специалистов других наук. СМИ твердят о том, что алгоритмы машинного обучения изобрели совсем недавно, буквально «на прошлой неделе», а данные никогда не были «большими», пока не появилась Google. Это совершенно не так. Множество современных методов и технологий (а также современных проблем) — частичная эволюция происходящего ранее. Это не значит, что нет ничего нового, достойного восхищения, но нам кажется важным указать на некоторые достижения прошлого, вызывающие уважение.
3. «Бесполезная шумиха», — бросают люди устало. Так описывают специалистов, работающих в области науки о данных, и это не сулит ничего хорошего. Чем больше шума, тем больше отвернувшихся и тем труднее разглядеть — есть что-то хорошее или его нет вовсе.
4. Статистики (ученые в области математической статистики) уверены, что уже работают в области «науки о данных». Это их хлеб с маслом. Может быть, вы, дорогой читатель, не статистик и вообще не имеете никакого отношения к этой области; просто представьте, что статистики воспринимают развитие отдельной науки о данных как покушение на их идентификацию, так же как это восприняли бы вы. Но мы намерены показать, что Data Science — не просто ребрендинг статистики или технологий машинного обучения, а, наоборот, самостоятельная наука. Мы хотим сделать это в противовес СМИ, которые часто описывают науку о данных просто как статистические методы или машинное обучение, применяемые в индустриальных технологиях.
5. Существует утверждение: «Нечто, вынужденное само себя называть наукой, таковой не является». В известной степени это так. *Сам по себе* термин «наука о данных» ничего не значит. А то, что скрывается под ним, не наука, а скорее искусство.

За пологом шумихи

Трудясь над получением степени PhD, Рэйчел приобрела значительный опыт в обработке статистических данных Google. Именно он иллюстрирует, несмотря на все вышеизложенное, те причины, по которым у нас появились подозрения, что в полемике, посвященной «науке о данных», возможно, есть здравое зерно. Вот ее слова:

> «Довольно быстро мне стало ясно, что реальная работа в Google совершенно не похожа на то, чему учили в школе. Я не хочу сказать, будто все ранее полученные знания оказались бесполезны. Наоборот, все изученное в школе явилось твердым, совершенно необходимым фундаментом, позволяющим выполнять мою работу.
>
> Однако множеству навыков, которые потребовались при решении рабочих задач, в школе *совершенно* не обучали. Конечно, мой опыт несколько специфичен, поскольку я владею навыками компьютерной обработки статистических данных, программирования, визуализации информации и обладаю знаниями предметной области Google. Любому другому, являющемуся только специалистом в области компьютерных или общественных наук, гуманитарием либо физиком, придется восполнить имеющиеся пробелы в знаниях. Однако здесь важно то, что, имея различные по глубине и охвату знания, мы сумели объединиться для решения проблем, возникших при работе с данными».

Вот так и возникла вся эта история. Всем известна прописная истина: приступая к реальной работе сразу после учебы, мы ощущаем разрыв между тем, чему нас учили, и тем, что требуется на работе. Другими словами, мы сталкиваемся с различием между академической статистикой и статистикой на производстве.

Несколько замечаний по этому поводу.

- Разумеется, между наукой и производством есть разница. Но действительно ли ей надлежит быть? Почему множество учебных курсов должны быть оторваны от действительности?
- Даже в таком случае отрыв не обусловлен различием между академической статистикой и производственной. Главное достижение адептов науки о данных в том, что их методология создает процессы, позволяющие работать *с бóльшим объемом знаний*, чем процессы, создаваемые фундаментальными методами статистики и компьютерных наук. Именно такие процессы мы определяем как *процессы науки о данных* (подробнее — в главе 2).

Ну а всю шумиху можно подытожить правдивым заключением: появилось нечто новенькое. Однако есть риск дальнейшего отвержения этой хрупкой, зарождающейся идеи. Прежде всего из-за ее позиционирования как некого чудодейственного средства, что порождает совершенно нереалистичные ожидания и, несомненно, приведет к разочарованию.

Рэйчел поставила перед собой задачу понять этот культурный феномен — науку о данных — и то, как даталогию воспринимают другие. Она стала встречаться с работниками Google, начинающими бизнесменами, представителями технологических компаний, университетов, преимущественно занятых в сфере статистики.

На этих встречах начала формироваться новая картина видения. Окончательное понимание вылилось в составление курса для Колумбийского университета под

названием «Введение в науку о данных», опубликованного Кэти в ее блоге. Мы решили, что по окончании семестра наряду с наиболее активными студентами будем понимать, что все это значит на самом деле. И теперь с помощью книги надеемся привести к этому же пониманию множество людей.

Почему именно сейчас

Мы накопили значительный объем данных о разных аспектах нашей жизни и в то же время имеем обилие недорогих компьютерных мощностей. Шопинг, общение, чтение новостей, прослушивание музыки, поиск информации — все это происходит онлайн и знакомо большинству людей.

А вот чего люди могут не знать: одновременно началась «датафикация» нашего поведении офлайн, и это обратная сторона онлайн-сбора информации (подробнее — ниже). Достаточно сложить два и два, чтобы получить четкое представление о нашем поведении и даже больше — о том, что мы вообще за вид.

Это касается не только интернет-данных, но и финансов, медицины, фармацевтики, социологии, правительства, образования, недвижимости; список можно продолжить. Значимость информации растет в большинстве секторов промышленности. В одних случаях объемы собираемой информации достаточно велики, чтобы называть их «большими» (подробнее — в главе 2), в других случаях это не так.

Но не только объемность этих новых данных делает их интересными (или создает проблемы). Данные, часто в режиме реального времени, служат строительными блоками для создания *информационного продукта*. В Интернете это рекомендательная система Amazon, рекомендации друзей в Facebook, советы по поводу фильмов, музыки и т. д. В финансах это представлено кредитными рейтингами, торговыми стратегиями и моделями. В образовании это приводит к пониманию динамики персонального обучения и оценки, проводимых, например, в Академии Хана. Для правительства это политика, основанная на информации.

Мы являемся свидетелями начала огромного цикла, насыщенного обратными связями, в котором наше поведение изменяет конечный продукт, а продукт изменяет наше поведение. Технологии делают это возможным: создается инфраструктура для крупномасштабной обработки данных, увеличиваются память и пропускная способность, а также культура использования технологий в построении нашей жизни. Еще десять лет назад это было невозможно.

Учитывая влияние обратной связи на весь цикл, следует серьезно подумать о том, как он проводится, а также об этических и технических обязательствах людей, ответственных за данный процесс. И первейшая цель нашей книги — начать именно этот разговор.

Датафикация. В мае — июне 2013 года издательство Foreign Affairs опубликовало статью *The Rise of Big Data* («Возникновение больших данных») Кеннета Нейла Кюкера (Kenneth Neil Cukier) и Виктора Майер-Шенбергера (Viktor Mayer-Schoenberger) (https://www.foreignaffairs.com/articles/2013-04-03/rise-big-data). В ней обсуждается концепция датафикации на примере того, как мы оцениваем дружественный контент с помощью лайков. Конечный вывод таков: все, что мы делаем онлайн, заканчивается записью в чье-либо хранилище данных для последующего изучения или продажи.

Авторы определят датафикацию как процесс «представления всех аспектов жизни в виде данных». Примером датафикации видения служат очки дополненной реальности компании Google. Twitter — образец датафикации случайных мыслей. LinkedIn — сеть для датафикации профессионализма.

Датафикация — очень интересная концепция, она вынудила нас признать ее важность в плане совместного использования информации разными людьми. Нас — или скорее наши действия, когда мы ставим лайки или запрашиваем какие-либо сведения, — датафицируют. Ну или как минимум мы должны быть готовы к этому. Едва мы открываем браузер, тут же (пусть непреднамеренно и неосознанно) датафицируемся с помощью cookie-файлов, о которых можем и не знать. Даже когда мы гуляем по магазину или просто по улице, мы непреднамеренно датафицируемся через сенсорные датчики, камеры наблюдения или очки Google.

Уровень преднамеренности колеблется в очень широком спектре: от восторженного участия в экспериментах, проводимых в социальных сетях, чем мы порой гордимся, до скрытого наблюдения и преследования. Но это все — датафикация. Наши намерения различны, но результат один.

В статье обращает на себя внимание строка, в которой авторы говорят о перспективах ценностей:

> «В результате датафикации вещей изменится их значимость, а информация превратится в новую форму ценностей».

Возникает важный вопрос, который мы будем поднимать на протяжении всей книги: кто «мы» есть в таком случае? Какого типа *ценности* имеются в виду? Большая часть приведенных примеров говорит о том, что «мы» — это модели собственников, которые зарабатывают деньги, побуждая людей покупать их вещи. Тогда под ценностью понимается нечто увеличивающее эффективность продаж с помощью автоматизации процесса.

Но если мы хотим мыслить шире, если хотим говорить о «нас» как людях вообще, то придется плыть против течения.

Сегодняшняя картина (и немного истории)

Итак, что такое наука о данных? Нечто новенькое или ребрендинг статистики и аналитики? Есть в ней зерно или это просто шумиха? И если это нечто новое и оно настоящее, то в чем оно заключается?

Это давняя дискуссия, и единственный способ понять, что происходит в данной индустрии, — обратиться к Интернету и посмотреть, в какой стадии находится обсуждение. Это не обязательно даст представление о даталогии, но зато мы увидим, что думают люди по данному поводу и как ее воспринимают. Так, на Quora[1] обсуждение *What is Data Science* («Что такое наука о данных?») длится с 2010 года. А вот пример ответа на этот вопрос, который дает генеральный директор Metamarket Майк Дрисколл (Mike Driscoll) (https://www.quora.com/What-is-data-science):

> «Практически наука о данных — это смесь хакерства, которое заряжается пивом Red Bull, и статистики, вдохновляющейся кофе эспрессо.
>
> Однако Data Science — не совсем хакерство ввиду того, что когда хакеры отлаживают свои однотипные приложения типа Bash или Pig, едва ли их волнует неевклидова геометрия.
>
> Но это и не совсем статистика, ведь когда статистики заканчивают теоретизировать какую-либо модель, едва ли кто-то из них способен прочесть файл на языке R, который составлен по результатам их работы.
>
> Даталогия — гражданское проектирование данных. Ее адепты обладают не только практическими знаниями инструментов и материалов, но и теоретическим пониманием того, что возможно».

Далее Дрисколл ссылается на диаграмму Венна о науке о данных, разработанную Дрю Конвеем (Drew Conway) в 2010 году (http://drewconway.com/zia/?p=2378) и приведенную на рис. 1.1.

Он также упоминает статью Натана Яу (Nathan Yau) *Rise of the Data Scientist* («Возникновение науки о данных»), опубликованную в 2009 году, и перечисляет замечательные навыки фанатов науки о данных в таких областях, как:

- статистика (традиционный анализ, о котором вы привыкли думать);
- очистка данных (разбор, очистка и форматирование);
- визуализация (графики, инструменты и т. д.).

Но позвольте, значит, наука о данных — просто набор специфических приемов? Или это логическое расширение других областей, таких как статистика и машинное обучение?

---

[1] Популярный калифорнийский сайт. — *Примеч. пер.*