Разработка → Что такое большие данные, часть 3 перевод

Data Mining*, Big Data*, Amazon Web Services*
Источник: https://habrahabr.ru/post/311460/

В первой части мы узнали о данных, и о том, как они могут быть использованы для извлечения из них метаданных или каких-то значений.

Вторая часть объяснила сам термин Big Data и показала, как он превратился в индустрию, причиной появления для которой стало влияние экономики. Эта, третья часть, в которой должно быть логическое продолжение предыдущих двух и у всего этого должен появиться смысл — грустная, местами ироничная, а местами пугающая. Вы видите сами, как технологические, бизнес, и даже социальные контракты в перспективе уже переопределялись большими данными таким путём, который мы только сейчас начинаем понимать. И, возможно, они никогда уже не станут контролируемыми.

С помощью чего бы не проводился анализ — суперкомпьютера или составленной вручную в 1665 году таблицы из списков мёртвых, некоторые аспекты больших данных существовали гораздо дольше, чем мы можем представить.

Темная сторона больших данных. Исторически роль больших данных не всегда была кристально чистотой. Идея переработки цифр, приводящей к количественной рационализации для чего-то, что мы и так хотели сделать, существует с тех пор, как у нас появились лишние деньги.

Помните корпоративных рейдеров из 80-х и их новомодное оружие — электронные таблицы? Электронная таблица — рудиментарная база данных, позволяла 27-летнему бакалавру с ПК и тремя обрывками сомнительных данных вовлекать своих начальников в грабёж из пенсионного фонда компании и выкупать контрольные пакеты акций за счет кредита. Без персональных компьютеров и электронных таблиц не было бы ни Майклов Милкенсов (Michael Milkens), ни Эйванов Боески (Ivan Boeskys). Это была всего лишь версия больших данных периода 80-х. Педанты скажут, что это не большие данные, но в культурном плане они имели тот же эффект, какой имеет индустрия, которую сегодня мы называем большими данными. Для того времени это были большие данные.

Помните рейганомику? Экономист Артур Лаффер утверждал, что поднять доходы государства можно за счет снижения налогов для богатых. Некоторые люди до сих пор в это верят, но они ошибаются.

Программный трейдинг — покупка и продажа акций с помощью компьютерного алгоритма развалил Уолл-стрит в 1987 году, потому что фирмы приняли его, но до конца не понимали, как использовать. Каждый отдельный компьютер не понимал, что другие компьютеры работают одновременно и, возможно, в ответ на те же правила, превращая то, что должно было быть организованным отступлением в паническую распродажу.

Менеджмент долгосрочного капитала в 90-х поставил вторичные ценные бумаги в такое положение, в котором они раньше никогда не были, только для того, чтобы феерически провалиться, потому что никто не понимал, что такое вторичные ценные бумаги. Если бы правительство вовремя не вмешалось, Уолл-стрит снова бы обрушился.

Enron в начале 2000-х годов использовал гигантские компьютеры для игры на энергетических рынках, или им так казалось, пока компания не схлопнулась. История компании Enron, помните, базировалась на том, что вычислительные гиганты делали компанию умнее, а в реальности использовались, чтобы маскировать обман и манипулировать рынком. "Не обращайте внимания на человека за занавеской!"

Мировой банковский кризис 2007 года частично был спровоцирован крупными компьютерами, для создания предположительно совершенных финансовых продуктов, к которым в итоге рынок не смог адаптироваться. Но разве всё это не было вызвано дерегуляцией (уменьшением гос влияния на экономику)? Дерегуляция может склонять финансистов к безрассудству, но более важную роль сыграл закон Мура, и стоимость вычислений снизилась до уровня, когда стало вполне возможным технически насиловать регулирование.**Технология породила искушение погнаться за крупной добычей.**

Всё это просто вариации тёмной стороны больших данных. Схемы больших данных очень быстро и сильно раздулись, а потом рухнули. Эти закрученные данными безумства, вообще-то, должны быть основаны не на реальности, а на фантазии, которой можно как-то покрыть реальность.

И внутри такие неудачи, обычно, основываются на лжи или подрываются ложью. Как ещё можно сделать высокорейтинговые облигации, обеспеченные ипотекой используя только мусорные ипотеки? Врать.

Строить выводы на основе ошибочного метода или ошибочных данных — большая проблема. Некоторые эксперты поспорили бы, что это легко исправить с помощью более крупных больших данных. И, может быть, такое возможно, но предыстория подобных действий показывает нехорошие результаты.

Тут есть ирония, и она в том, что когда мы верим ошибочным большим данным, они обычно как-то связаны с деньгами или каким-то проявлением силы, но у нас такая же тенденция не верить верным большим данным, когда они касаются политики или религии. Поэтому большие научные данные часто борются за признание своих идей скептиками, теми, кто отрицает изменение климата или поддерживает обучение креационизму.

Большие Данные и страхование. Большие данные уже изменили наш мир во многих отношениях. Возьмите, например, медицинскую страховку. Было время, когда актуарии страховых компаний изучали статистику заболеваемости и смертности для установления страховых тарифов. Сюда были вовлечены метаданные — данные о данных — потому в основном актуарии не могли копнуть достаточно глубоко, чтобы добраться за пределы широких групп держателей страховых полисов физическими лицами. В такой системе рентабельность страховой компании линейно возрастает с ростом объёма клиентов. Прибыль с большинства клиентов небольшая, поэтому медицинские страховые компании нуждались в увеличении количества застрахованных — чем больше, тем лучше.

Потом в 90-е кое-что случилось: стоимость вычислений достигла уровня, когда стало экономически эффективным рассчитывать вероятные результаты о здоровье на индивидуальной основе. Это кинуло бизнес медицинского страхования от установления норм к отказу от компенсаций. В США бизнес-модель медицинского страхования перешла от охвата максимального количества людей к минимальному количеству и продаже страховки только здоровым людям, тем, кто не нуждается в здравоохранении.

Прибыли страховых компаний взлетели, но у нас появились миллионы незастрахованных семей.

Принимая во внимание, что обществу нужны люди здоровые, продажа страховок только здоровым, очевидно, не могла долго продолжаться. Возник новый вид экономического пузыря, который ожидал своего конца — поэтому появилась Obamacare (система здравоохранения с защитой пациента и доступностью обслуживания). На эту тему можно целую книгу написать, но нужно понимать, что что-то должно было произойти, чтобы система страхования изменилась, для достижения социальной цели.*Доверяя электронным умным системам каким-то образом находить характерный способ покрывать больший процент населения, продолжая при этом расширять границы прибыли — невменяемое поведение.*

Большой пугающий Google. Лохотрон, который слишком часто лежит в основе больших данных, проникает через всю экономику и затрагивает тех людей, которых мы решили считать воплощением больших данных или даже их создателями. Google, например, хочет, чтобы мы верили, что он знает, что делает. Я не говорю, что творения Google не фантастические и не важные, но они построены на недосягаемом алгоритме, который они не будут объяснять, так же как Берни Мэдофф не будет объяснять свои свои инвестиционные методы. Большой пугающий Google, он постиг вселенскую мудрость.

Возможно, но кто знает точно?

Правда в том, что все деньги делают они!

Правда о рекламе. Google выставляет счета рекламодателям, но весь их доход — это рекламодатели, платящие по счетам. Хотя это может показаться банальным, часто рекламодатели не слишком хотят знать, насколько хорошо идут дела у их компании. Если бы было достаточно прозрачно, насколько до смешного низкая доходность у большего объёма рекламы, рекламные агентства уходили бы из бизнеса. А поскольку агентства не только производят рекламу, а ещё и размещают её где-то для клиентов, с точки зрения рекламной индустрии иногда лучше не знать.

В результате среди рекламщиков структура власти перевернута с ног на голову. Вышестоящие работники агентства зарабатывают большую часть денег и входят в творческую прослойку, которая занимается непосредственно рекламой. Новички, которые почти не зарабатывают денег и почти не имеют авторитета — это те, кто публикуют печатную, медийную и даже интернет-рекламу. Рекламная индустрия оценивает создание рекламы дороже любой окупаемости для клиентов. Это безумие и чтобы такая схема работала, невежество должно торжествовать.

Получается, что интернет коррумпирован. Новостной агрегатор Huffington Post сказал своим авторам использовать термины поисковой оптимизации в публикациях, для предположительного увеличения читательской аудитории. Работает ли это? Не очень

ясно, хотя исследования говорят, что показатели оптимизатора поисковых систем идут вверх, если вставлять тарабарщину в посты, что вообще не имеет смысла.

В итоге, происходит следующее — мы поддаёмся и миримся с более низким стандартом производительности. Помогут ли вам Match.com или eHarmony найти лучшую пару? Нет. Но забавно думать, что они способны на это, так какого чёрта...

Опять же, странно здесь то, что мы цинично относимся к данным, когда речь идет о науке (изменении климата, креационизме и т.д.), но почти совсем без цинизма, когда речь идет о бизнес-данных.

Теперь интересно принять во внимание вот какой факт. Данные компании Google — необработанные, не проанализированные — в большем объёме доступные другим компаниям, так почему у Google нет эффективного конкурента в поиске? Bing от Microsoft, безусловно, имеет доступ к тем же данным, что и Google, но у них одна шестая пользователей. Здесь всё сводится к восприятию рынка, а Bing не воспринимается как подходящая альтернатива Google, хотя он именно такой.

Это игра больших данных.

Другая, неосвещённая, сторона этой истории. У Apple центр данных в Северной Каролине оценивается в $1 млрд, он построен ещё до смерти Стива Джобса. Я провел день, припарковавшись перед воротами этого объекта, и насчитал одну въезжающую и выезжающую машину. Я продолжал рассчитывать какая потребовалась бы серверная мощность, если бы Apple хранил в этом здании несколько копий всех существующих на Земле данных и насчитал восемь процентов доступного им на данный момент пространства. Это здание способно содержать два миллиона серверов.

Позже я встретился с торговым агентом, который продал компании Apple каждый сервер в этом здании — все 20 000, как он сказал.

Двадцать тысяч серверов много для iTunes, но они занимают один процент площади всего здания. Что там происходит? Это обман больших данных: потратив $1 млрд на строительство, Apple заглядывается на Уолл-стрит (и на конкурентов Apple), как игрок в игре Google.

Нельзя сказать, что большие данные нереальные, потому что они как раз реальные. У Amazon.com и любой другой огромной компании розничной торговли, включая Walmart, большие данные вполне реальны, потому что этим компаниям реальные данные нужны, чтобы быть успешными на тонких гранях рентабельности. Успех Walmart всегда строился на информационных технологиях. В электронной коммерции, где покупаются и продаются реальные вещи, клиент всегда остаётся клиентом.

Для Google и Facebook клиент — это продукт. Google и Facebook торгуют нами.

Все это время закон Мура успешно действует, наколдовывая всё более дешёвые и мощные вычисления. Как мы уже говорили в первой части, каждое десятилетие вычислительная мощность с той же ценой повышается на коэффициент 100, только благодаря закону Мура. Компьютерная транзакция необходимая для продажи авиабилетов через систему SABRE в 1955 году снизилась в миллиард раз к сегодняшнему дню. То, что было резонным расходом в $10 за билет в 1955 году, сегодня ничтожная часть пенни, которую нет смысла учитывать. В ценностной системе SABRE вычисления сегодня фактически бесплатные. Это полностью меняет то, что мы можем делать с помощью компьютеров.

Ваша личная служба разведки. Вычисления стали настолько недорогими, а персональные данные проникли настолько глубоко, что теперь некоторые облачные приложения превращают ваш смартфон в подобие машин для интеллектуального анализа данных Эдгара Гувера (ФБР) или NSA сегодня. Один из таких инструментов был назван Refresh и изображён на картинке. Refresh затем был поглощён LinkedIn, а тот поглощён Microsoft, но образец всё ещё действует. Введите чьё-нибудь имя в телефон и сотни компьютеров — в буквальном смысле сотни — обшарят социальные медиа и веб, составляя оперативное досье на человека, с которым у вас деловая встреча или вы просто хотите посидеть с ним в баре, и вы не просто видите всё об этом человеке: его жизни, работе, семье, образовании, система может отследить как пересекались ваши с ним жизни, предугадать вопросы, которые вы могли бы задать или темы разговора, которые вы, возможно, захотите развить. Все это в пределах одной секунды. И бесплатно.

Ну хорошо, цифровая шпаргалка — вряд ли, вершина развития компьютеризации, созданной человеком, но она показывает, как далеко мы продвинулись, и предполагает, насколько ещё сможем, поскольку вычисления становятся еще дешевле. И вычисления станут ещё дешевле, так как закон Мура не замедляется, а напротив — ускоряется.

Провал искусственного интеллекта. Еще в 80-х была популярна область, называемая искусственным интеллектом, основной идеей которой было выяснить, как эксперты делают то, что делают, свести эти задачи к набору правил, потом запрограммировать компьютеры используя эти правила и эффективно заменить экспертов. Цель состояла в том, чтобы научить компьютеры диагностировать заболевания, переводить языки, даже выяснять, чего мы хотим, но сами понять не способны.

Это не сработало.

Искусственный интеллект или, как его назвали, AI (Artificial Intelligence), насосал сотни миллионов венчурных долларов силиконовой долины, прежде чем был объявлен банкротом. Хотя в то время проблема искусственного интеллекта чётко не прорисовывалась, она заключалась в том, что нам просто не хватало вычислительной мощности по соответствующей времени цене для достижения этих амбициозных целей. Но благодаря Map Reduce и облачной инфраструктуре сегодня у нас есть более чем достаточно вычислительной мощности, чтобы создать искусственный интеллект.

Лежачий полицейский Парадоксально то, что ключевой идеей искусственного интеллекта было дать язык компьютерам, а в реальности случилось такое, что значительная часть успеха компании Google оказалась в эффективном отдалении языка от компьютеров, человеческого языка. Стандарты данных XML и SQL, которые лежат в основе почти всего веб-контента, не используются в Google, потому что они поняли, что структуры данных, приспособленные к чтению человеком не имеют смысла для компьютеров, которые будут общаться между собой. За счет того, что человек больше не требовался для компьютерной коммуникации, был достигнут значительный прогресс в области машинного обучения. **Это очень важно, пожалуйста, прочитайте это ещё раз.**

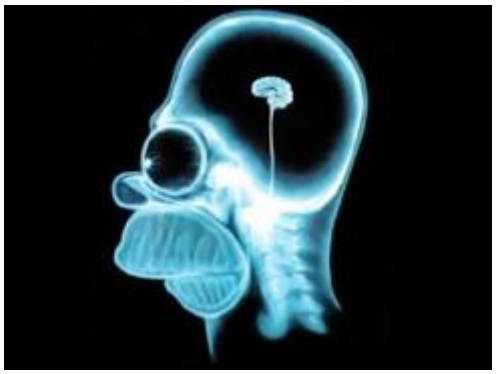
Понимаете, в современной версии искусственного интеллекта нам не нужно обучать компьютеры выполнению человеческих задач: они сами себя учат.

Google Translate, например, может использовать онлайн, бесплатно кто угодно, для перевода текста в разных комбинациях между более чем 70 языками. Этот статистический переводчик использует миллиарды последовательностей слов, которые отображаются на двух или более языках. Вот это на английском означает вот это по-французски. Никаких частей речи, никаких подлежащих или глаголов, никакой грамматики вообще. Система просто выясняет это. А значит, что для теории нет никакой необходимости. Это четко работает, но мы не можем точно сказать — как, потому что весь процесс управляется данными. Со временем Google Translate будет совершенствоваться всё больше, делая перевод на основе так называемых корреляционных алгоритмов — правил, которые никогда не покидают машину и слишком сложны для людей, чтобы даже понять.

Мозг Google. В Google есть одна штука, которая называется Google Vision, совсем недавно в ней было 16 000 микропроцессоров — эквивалент примерно одной десятой зрительной коры человеческого мозга. Он специализируется на компьютерном зрении и обучается таким же образом, как Google Translate, с помощью огромного количества образцов — в данном случае неподвижных изображений (миллиарда неподвижных изображений), которые берутся из видео на YouTube. Google Vision рассматривает изображения 72 часа и, по сути, обучает себя распознавать в два раза больше, чем любой компьютер на Земле. Дайте ему изображение, и он найдет еще одно подобное. Скажите ему, что на изображении кошка, и он будет способен распознавать кошек. Помните, на это уходит три дня. Сколько времени занимает распознать кошку у новорожденного ребёнка?

Точно так же, как Уотсон из IBM выиграл в Jeopardy (российская версия телепередачи называется «Своя Игра», — прим. пер.), просто перерабатывая вопросы из прошлых выпусков: не было никакой лежащей в основе теории.

Давайте продвинемся еще на пару шагов. Проводили исследования, управляемые данными на основе магнитно-резонансных томограмм (МРТ), изображений живых мозгов осужденных преступников. Эта система не отличается от примера с Google Vision, кроме того, что мы разбираем тут другой вопрос — рецидивизм, вероятность того, что преступник нарушит закон снова и вернётся в тюрьму после освобождения. Опять же, без какой-либо базовой теории Google Vision, кажется, способен различать те МРТ-снимки уголовников, которые способны на повторное преступление и те, которые не способны. Гугловский коэффициент результативности для прогнозирования совершения преступления, основывается исключительно на одном скане мозга и составляет 90+ процентов. Должны ли снимки МРТ стать инструментом для принятия решения каких заключенных освобождать условно-досрочно? Звучит немного похожим на фильм Minority Report («Особое мнение») с Томом Крузом. В этой схеме есть огромная предположительная экономическая выгода для всего общества, но она содержит страшный аспект отсутствия теории, лежащей в основе: она работает, потому что работает.



После этого ученые Google посмотрели на МРТ обычных людей в то время как те просматривали миллиарды кадров YouTube. Переработав достаточно большой набор данных этих изображений и результирующие МРТ, компьютер может предсказать, на что смотрит субъект.

Это называется чтением мыслей… и, снова, мы не знаем, как это работает.

Продвигаем науку, устраняя ученых.

Что делают ученые? Они теоретизируют. Большие данные в некоторых случаях делают теорию ненужной или просто невозможной. В 2013 году Нобелевская премия в химии была присуждена тройке биологов, которые все свои исследования построили на выводе, сделанном компьютерными алгоритмами, чтобы объяснить химию энзимов. При получении этой премии ни один энзим не пострадал.

Алгоритмы сегодня совершенствуются в два раза быстрее закона Мура.

Что меняется, так это возникновение нового рабочего процесса информационной технологии, который начинается с традиционного:

1. Новое железо порождает новый софт.
2. Новый софт пишется для новых областей деятельности, обеспеченных новым железом.
3. Закон Мура снижает стоимость железа со временем, и новый софт становится ориентированным на потребителя.
4. Смыть, повторить.

К следующему поколению:

1. Массовый параллелизм позволяет органично получать новые алгоритмы
2. Новые алгоритмы работают на потребительской технике
3. Закон Мура эффективно форсируется, хотя и с некоторыми рисками (мы не понимаем свои алгоритмы)
4. Смыть, повторить

В чём тут суть? Новый стиль внедрения выходит за рамки того, что всегда требуется для значительного технологического скачка — новой вычислительной платформы. Что будет после мобильных телефонов, спрашивают люди? Это и будет после мобильных телефонов. Как это будет выглядеть? Никто не знает, и возможно всё это будет не важно.

Через 10 лет закон Мура увеличит мощность процессора в 128 раз. Натравливая больше процессорных ядер на решение задач и эксплуатируя быстрыми темпами развития алгоритмов, мы должны увеличить это значение ещё в 128 раз: в общей сложности в 16,384. Помните, Google Vision в настоящее время — это эквивалент 0,1 объёма зрительной коры. Теперь умножьте это на 16 384 и получите 1 638 эквивалентов зрительной коры. Вот к чему это ведет.

Через десять лет компьютерное зрение сможет видеть вещи, которые мы не понимаем, так же, как собаки могут унюхать рак.

Мы бьёмся о стену нашей способности генерировать соответствующие теории, одновременно находя в больших данных хаки, чтобы любыми способами продолжать улучшать результаты. Единственная проблема только в том, что мы больше не понимаем, как что-то работает. Как много времени осталось до момента, когда мы полностью потеряем контроль?

Примерно к 2029, согласно Рэю Курцвейлу, мы достигнем технологической сингулярности.

В этом году говорит знаменитый футурист (и гуглер), за $1 000 можно будет приобрести вычислительную мощность, которая будет соответствовать 10 000 человеческим интеллектам. За цену ПК, говорит Рэй, мы сможем использовать больше вычислительной мощности, чем мы можем понять или даже объяснить. Настоящий суперкомпьютер в каждом гараже.

В сочетании со столь же быстрыми сетями это может означать, что ваш компьютер — или какое бы устройство у вас не было — может обшарить в реальном времени абсолютно все когда-либо написанные слова, чтобы ответить буквально на любой заданный вопрос. Не оставляя ни одного не обитого порога.

Спрятаться не выйдет. Примените это к миру, где каждое электрическое устройство — это датчик подающий сигнал сети, и у нас будут не только невероятно эффективные пожарные сигнализации, мы скорее всего потеряем любую приватность.

Те, кто предсказывает будущее склонны переоценивать изменения на краткий срок и недооценивать долгосрочные. Desk Set 1957 с Кэтрин Хепберн и Спенсером Трейси предвидели автоматизацию на базе мэйнфреймов и исключения людей, управляющих машинами в научно-исследовательском отделе телесетей. В какой-то степени это сбылось, хотя потребовалось еще 50 лет и люди остались частью процесса. Но самая крупная технологическая угроза висела не над научно-исследовательским отделом, а над самой телевизионной сетью. Будут ли существовать телевизионные сети в 2029 году? Будет ли существовать телевидение вообще?

Никто не знает.

Если вы прочитали всю серию статей и случайно оказались работником Google, вы можете чувствовать, что вас атаковали, потому что многое из того, что я описываю, может угрожать текущему образу жизни, а имя "Google" часто встречается в тексте. Но это не так. Точнее, это не совсем так. Google — удобная мишень, но такую же работу выполняют прямо сейчас компании вроде Amazon, Facebook и Microsoft, и ещё около ста или может больше других стартапов. Google — не единственный. И регулирование деятельности Google (что пытаются делать европейцы) или попытки выкинуть его из бизнеса, вероятно, ничего не изменят. Будущее наступает не смотря ни на что. Пять из этих сотен стартапов будут иметь феерический успех, и их будет достаточно, чтобы навсегда изменить мир.

Так мы пришли к самоуправляемому автомобилю. Такие компании, как Google и его конкуренты процветают благодаря производству всё более быстрых и дешёвых вычислений, потому что это делает их вероятными поставщиками продуктов и услуг, управляемых данными, в будущем. Это будущее индустрии.

Сегодня, если взять стоимость деталей современного автомобиля, то пучок проводов, который соединяет все электрические биты и контролирует весь механизм, стоит больше, чем двигатель и коробка передач! Это показывает какой у нас приоритет: команды и коммуникации, а не движение. Но эти затраты резко снижаются, а их функциональные возможности увеличиваются с той же скоростью. Сумма в $10 000,

выделенная Google для самоуправляемого авто, упадёт до нуля через десятилетие, после того, как все новые автомобили станут самоуправляемыми.

Сделайте все новые автомобили самоуправляемыми и природа автомобильной культуры полностью изменится. Автомобили появятся везде, они будут ехать с максимальной разрешенной скоростью, и между ними будет всего один метр. Это увеличит пропускную способность автомобильных дорог в 10 раз.

Тот же эффект может сказаться на авиаперелётах. Самоуправляемые самолёты могут привести к появлению большого количества мелких самолетов, которые будут как стаи птиц лететь прямо в место назначения.

А может быть, мы вообще перестанем путешествовать. Увеличенные вычислительные мощности и более быстрые сети уже дают возможность телеприсутствия — видеоконференцсвязь в натуральную величину, там, где нужно потребителю.

Возможно единственным реальным общением с людьми за пределами своей деревни, будут моменты, когда мы будем к ним физически прикасаться.

Всё это и многое другое вероятно. Биоинформатика — применение массивной вычислительной мощности в медицине, в сочетании с корреляционными алгоритмами и машинным обучением, будут искать ответы на вопросы, которых мы еще не задали и никогда не зададим.

Возможно, мы победим и заболевания и старение, а значит, что умирать мы будем от рук преступников, самоубийств, или трагических несчастных случаев.

Компании с большими данными несутся сломя голову, захватывая важные позиции поставщиков будущего. Закон Мура далеко вышел за грань, где этом стало неизбежным. Мы дошли до точки невозврата.

И мир полностью меняется, а мы можем только догадываться, как он будет выглядеть и кто… или что… будет его контролировать.

(Перевод Наталии Басс)