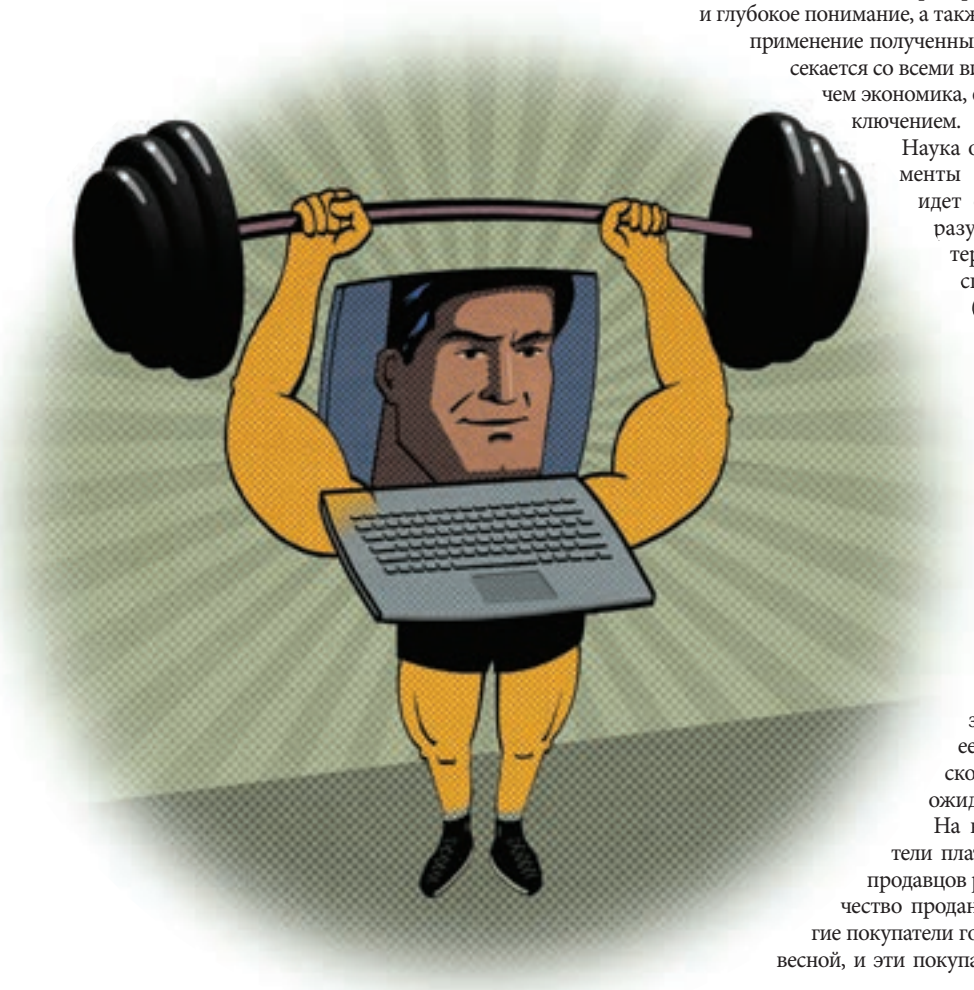
Большая сила больших данных

Вычислительные мощности дают импульс машинному самообучению и ведут к преобразованию бизнеса и финансов

Саджив Раджан Дас

СЕГОДНЯ миру доступен гораздо больший объем данных, чем можно было представить себе всего десять лет тому назад. Предприятия накапливают новые данные быстрее, чем могут их организовать и осмыслить. Теперь им приходится задуматься о том, как использовать этот огромный объем данных, чтобы принимать более обоснованные решения и повышать свою результативность.

Новая область науки о данных призвана обеспечить извлечение информации, обладающей практической ценностью, из данных, особенно из больших данных — чрезвычайно крупных массивов данных, анализ которых может привести к выявлению закономерностей, тенденций и взаимосвязей. Наука о данных охватывает как сбор и организацию данных, так и их анализ и глубокое понимание, а также, в конечном итоге, практическое применение полученных знаний. Эта область науки пересекается со всеми видами деятельности человека, причем экономика, финансы и бизнес не являются исключением.



Наука о данных ввела в обиход инструменты машинного самообучения: речь идет об особом виде искусственного разума, который наделяет компьютеры способностью к обучению без специального программирования (Samuel, 1959). Эти инструменты, в сочетании с обширными объемами данных, способны радикальным образом изменить ситуацию в области управления предприятиями и анализа экономической политики.

Некоторые из этих изменений выглядят весьма многообещающими.

Профилирование потребителей

Учитывая убедительные экономические аргументы в пользу науки о данных, быстрый рост ее применения в предпринимательской деятельности нельзя назвать неожиданным.

На конкурентном рынке все покупатели платят одну и ту же цену, а выручка продавцов равна цене, помноженной на количество проданных товаров. Тем не менее, многие покупатели готовы заплатить цену выше равновесной, и эти покупатели сохраняют потребительский

излишек, который можно извлечь, используя для профилирования потребителей большие данные.

Взимая с потребителей разную цену в зависимости от их профиля, полученного в результате анализа, компании получают возможность выручить самую высокую цену, которую потребители хотят и могут платить за тот или иной конкретный товар. Оптимизация ценовой дискриминации или сегментация рынка с помощью больших данных приносит очень большую выгоду. Эта практика уже являлась нормой в некоторых отраслях (например, в сфере авиаперевозок), но в настоящее время она распространяется на все виды продукции.

Кроме того, прибыль, полученная благодаря ценовой дискриминации, позволяет компаниям предоставлять скидки потребителям, которые в ином случае не смогли бы позволить себе заплатить равновесную цену, и это ведет к увеличению доходов и расширению клиентской базы, а также, возможно, к росту общественного благосостояния. Профилирование потребителей с использованием больших данных — одна из важных причин высокой стоимостной оценки таких компаний, как Facebook, Google и Acxiom, предлагающих товары и услуги на основе данных их клиентов.

Хотя большие данные могут использоваться для получения выгоды за счет потребителей, они в то же время меняют предпринимательскую практику на благо тех же потребителей. Компании используют данные, полученные путем анализа взаимодействия между людьми в социальных сетях, чтобы лучше понять их кредитное поведение. Увязывание предшествующей кредитной истории физических лиц с их присутствием в социальных сетях ведет к совершенствованию систем оценки кредитоспособности. Кроме того, оно позволяет кредитным организациям предоставлять кредиты тем, кому в противном случае могло бы быть отказано.

В частности, большие данные помогают устранять систематические ошибки, возникающие при принятии решений на основе ограниченной информации. Нехватка детализированных индивидуальных данных привела к дискриминации при оценке заявок на получение кредита в зависимости от городских районов (практика «красной черты»), которая ведет свое начало с 1930-х годов. Организации, предоставляющие ипотечные кредиты, обозначали красными линиями участки на карте, чтобы показать, что они не будут выдавать кредиты на недвижимость в этих районах из-за их расового или этнического состава. Эта практика, основанная на стереотипах, лишила доступа к кредитам целые сегменты общества.

Однако благодаря большим данным применение стереотипов уходит в прошлое. Отныне приблизительные субъективные данные могут быть заменены более точными данными, носящими более индивидуализированный характер. Компании, занимающиеся оценкой кредитоспособности, могут использовать все многообразие данных, выводимых из взаимодействия людей в социальных сетях, текстовых потоков, микроблогов, моделей использования кредитных карт и данных профилирования, в дополнение к таким обычным демографическим сведениям, как доход, возраст и место жительства (Wei et al., 2014). Использование более детализированных данных позволяет точнее оценивать физических лиц по их кредитным характеристикам.

Прогнозирование и анализ риска

Благодаря методам науки о данных в экономическом прогнозировании произошли радикальные перемены. В рамках традиционного прогнозирования важнейшая экономическая статистика, например, отчеты о квартальном ВВП, поступает со значительным опозданием. Наука о данных может решить проблему подобных задержек, так как использует информацию, поступающую с более высокой периодичностью, такую как данные о безработице, производственные заказы или даже новостной фон, для предсказания переменных, информация о которых поступает реже.

Набор подходов, применяющихся в этой деятельности, известен как «сверхкраткосрочное прогнозирование» (или «наукастинг»), но правильнее всего будет определить его как прогнозирование в режиме реального времени (см. статью «Царица цифр» в мартовском номере *Ф&Р* за 2014 год).

Наука о данных заявляет о себе и тогда, когда речь заходит об анализе системного финансового риска. Мир стал как никогда взаимосвязанным, и определение масштабов этих связей обещает новый уровень понимания ситуации для принятия экономических решений.

Подход, основанный на рассмотрении системного риска через призму сетей, обладает огромным потенциалом. Сегодня исследователи, специализирующиеся на анализе данных, используют многочисленные данные для того, чтобы воссоздать картину взаимодействия между банками, страховыми компаниями, брокерами и т. д. Безусловно, полезно знать, какие банки обладают более развитыми связями. Не менее полезна и информация о том, какие банки пользуются наибольшим влиянием, и это можно выяснить с использованием метода, основанного на характеристических значениях. Выстроив такие сети, специалисты по анализу данных могут оценить степень риска в той или иной финансовой системе, равно как и вклад отдельных финансовых организаций в общий риск, и это предлагает органам регулирования новый способ анализа системного риска и, в конечном итоге, управления таким риском. См. Espinosa-Vega and Solé (2010); МВФ (2010); Burdick et al. (2011) и Das (2016).

Эти подходы многое заимствуют из разработанных в социологии математических моделей социальных сетей и применяются в очень крупных сетях с использованием передовых вычислительных моделей, что в конечном итоге ведет к плодотворному синтезу целого ряда академических дисциплин.

Больше, чем слова

Аналитическая обработка текстов представляет собой быстро растущую область науки о данных и служит интересным дополнением к количественным данным в финансово-экономической сфере (см. статью «Две стороны перемен» в этом номере *Ф&Р*). Существует огромное количество вариантов коммерческого применения технологий, основанных на интеллектуальном анализе текста: такие компании, как iSentium, составляют долгосрочные и краткосрочные прогнозы настроений на основе социальных сетей, используя Twitter; a StockTwits предоставляет индикаторы настроений через веб-приложение, приспособленное для мобильных телефонов.

Сейчас можно определить рейтинг той или иной компании по показателям квартальных доходов, приведенным в ее ежегодном отчете о финансовых результатах по форме 10-K, который подается в Комиссию по ценным бумагам и биржам США. Определенный набор слов, отсылающих к риску, в ежеквартальных отчетах предоставляет в распоряжение аналитика точную систему ранжирования, позволяющую прогнозировать доходы. Компании, ежеквартальные отчеты которых труднее читать, обычно имеют более низкий доход. Скорее всего, это происходит потому, что они стараются сообщать плохие новости запутанным языком (см. Loughran and McDonald, 2014). Старый способ оценки удобочитаемости (индекс туманности Ганнига или фог-индекс) позволяет с легкостью оценивать финансовые отчеты по этому параметру, и такие органы регулирования, как Бюро по финансовой защите потребителей, подумывают о том, чтобы установить определенные стандарты удобочитаемости.

Исследования показали, что уже по одному лишь объему ежеквартальных отчетов можно распознать плохие новости (длинные отчеты предрекают сокращение доходов), и вновь это происходит потому, что запутанность изложения коррелирует с многословием. В конечном итоге это означает, что размер загруженного на веб-сайт Комиссии по ценным бумагам и биржам файла с документами компании сам по себе свидетельствует о показателях квартальных доходов. Ожидается, что эта стремительно развивающая область деятельности откроет еще больше новых возможностей.

Новый вид деятельности, известный как «новостной анализ», заключается в извлечении данных из новостных сообщений. Такие компании, как RavenPack, предоставляют все больше подобных услуг. Их диапазон включает в себя от оценки настроений и прогнозной аналитики для нужд торговли до макроэкономического прогнозирования. RavenPack извлекает информацию из огромного количества неструктурированных данных, содержащихся в новостях и социальных СМИ, и преобразует ее в детализированные данные и индикаторы, которые используются финансовыми компаниями для оценки активов, маркет-мейкерской деятельности, управления рисками и в обеспечении соблюдения норм.

В этой категории особый интерес представляет анализ потока новостей. Хеджевые фонды анализируют тысячи новостных лент в день, чтобы выделить пять или десять главных тем, а затем отслеживают динамику соотношения тем с течением времени, выявляя значимые для трейдинга изменения в конъюнктуре рынка. Подобный анализ был бы полезен директивным органам и регуляторам, таким как центральные банки. Например, резкое изменение в соотношении определенных тем, обсуждаемых в новостях (таких как инфляция, обменный курс или рост), может свидетельствовать о том, что пора пересмотреть политику процентных ставок.

Анализ тем начинается с построения гигантской таблицы частотности слов, которая известна как «терм-документная матрица» и позволяет каталогизировать тысячи новостных сообщений. Термины (слова) размещаются в горизонтальных рядах таблицы, тогда как каждой новостной статье отводится по колонке. Эта огромная матрица может обнаруживать темы посредством математического анализа корреляции между словами и между документами. Группы слов индексируются, а темы выявляются с помощью машинного самообучения, такого как латентно-семантическое индексирование и латентное размещение Дирихле (LDA). Анализ LDA позволяет получать наборы тем и списки слов, которые появляются в рамках этих тем.

Эти методы моделирования носят слишком узкоспециализированный характер, чтобы подробно рассматривать их здесь, но по сути они представляют собой лишь статистические технологии для обнаружения основных групп слов в определенном наборе документов (например, в потоке новостей). Эти языковые ключи могут широко использоваться органами, определяющими экономическую политику, а также при принятии политических решений, например, относительно новой формулировки политической программы в ходе той или иной политической кампании.

Искусственный разум и будущее

Сегодня компьютеры обладают большей мощностью, чем когда бы то ни было, а их способность обрабатывать большие объемы данных стимулировала прогресс в области искусственного разума. Новый вид алгоритмов, известных как «сети глубокого обучения» (термин отсылает к биологическим нейронным сетям), как выяснилось, обладает огромным потенциалом в имитации работы мозга, приведя к созданию множества успешных образцов искусственного разума.

Глубокое обучение представляет собой статистическую методику, которая использует искусственные нейронные сети для преобразования большого количества входных переменных в выходные переменные, то есть для выявления закономерностей. Информация подвергается анализу, проходя через сеть нейронов на кремниевой основе с программным обеспечением. Данные используются для укрепления связей между этими нейронами, что во многом аналогично обучению людей, извлекающих уроки из опыта с течением времени. У ошеломительного успеха глубокого обучения есть две причины: наличие огромных объемов данных, которые могут использоваться для обучения машин, и экспоненциальный рост вычислительной мощности, вызванный разработкой специальных компьютерных чипов для приложений, предназначенных для глубокого обучения.

Глубокое обучение лежит в основе многих современных технологий, которые мы с вами начинаем воспринимать как нечто само собой разумеющееся, будь то машинный перевод, автопилотируемые автомобили или распознавание и маркировка изображений. По всей вероятности, такие технологии могут очень скоро изменить экономику и политику. Агентства кредитного рейтинга уже используют их для подготовки отчетов без участия человека. Крупные нейронные сети глубокого обучения скоро смогут успешнее осуществлять прогнозы и выявлять связи между экономическими переменными, чем стандартные статистические методы.

Трудно предсказать, в каких областях «мрачной науки» использование машинного обучения будет расти особенно активно, но новая эра безусловно уже началась. Как сказал известный писатель-фантаст Уильям Гибсон, «будущее уже здесь, оно просто не совсем равномерно распределено». ■

Саджив Раджан Дас — профессор Школы бизнеса им. Ливи при университете Санта-Клары.

Литература:

Billio, Monica, Mila Getmansky, Andrew W. Lo, and Loriana Pelizzon, 2012, "Econometric Measures of Connectedness and Systemic Risk in the Finance and Insurance Sectors," Journal of Financial Economics, Vol. 104, No. 3, pp. 535–59.

Burdick, Douglas, Mauricio A. Hernandez, Howard Ho, Georgia Koutrika, Rajasekar Krishnamurthy, Lucian Popa, Ioana Stanoi, Shivakumar Vaithyanathan, and Sanjiv Das, 2011, "Extracting, Linking and Integrating Data from Public Sources: A Financial Case Study," IEEE Data Engineering Bulletin, Vol. 34, No. 3, pp. 60–7.

Das, Sanjiv, 2016, "Matrix Metrics: Network-Based Systemic Risk Scoring," Journal of Alternative Investments, Vol. 18, No. 4, pp. 33–51.

Espinosa-Vega, Marco A., and Juan Solé, 2010, "Cross-Border Financial Surveillance: A Network Perspective," IMF Working Paper 10/105 (Washington: International Monetary Fund).

International Monetary Fund (IMF), 2010, "Systemic Risk and the Redesign of Financial Regulation," Global Financial Stability Report, Chapter 2 (Washington, April).

Lin, Mingfen, Nagpurnanand Prabhala, and Siva Viswanathan, 2013, "Judging Borrowers by the Company They Keep: Friendship Networks and Information Asymmetry in Online Peer-to-Peer Lending," Management Science, Vol. 59, No. 1, pp. 17–35.

Loughran, Tim, and Bill McDonald, 2014, "Measuring Readability in Financial Disclosures," Journal of Finance, Vol. 69, No. 4, pp. 1643–71.

Samuel, A.L., 1959, "Some Studies in Machine Learning Using the Game of Checkers," IBM Journal of Research and Development, Vol. 3, No. 3, pp. 210–29.

Wei, Yanhao, Pinar Yildirim, Christophe Van den Bulte, and Chrysanthos Dellarocas, 2015, "Credit Scoring with Social Data," Marketing Science, Vol. 352, pp. 234–58.