УДК 004.855:519.216 https://doi.org/10.15407/pp2019.02.047

О.С. Балабанов

АНАЛІТИКА ВЕЛИКИХ ДАНИХ: ПРИНЦИПИ, НАПРЯМКИ І ЗАДАЧІ (ОГЛЯД)

Висвітлено основні напрямки, задачі та типи результатів глибокого аналізу великих (комп'ютеризованих) даних. Показано практичне значення великих даних та великої аналітики як фундаменту створення нових комп'ютерних технологій планування і керування у бізнесі. Виділено специфічні для великих даних режими використання даних (або роди завдань аналізу): «інтелектуальний» пошук потрібної інформації; масована переробка («відпрацювання») даних; індукція моделі об'єкту (середовища); екстракція знань з даних (відкриття структур і закономірностей). Окреслено етапи і організацію циклу робіт з аналізу даних. До типових класів задач великої аналітики належать: групування випадків (кластеризація); виведення ціле-визначених моделей (класифікація, регресія, розпізнавання); виведення генеративних моделей; відкриття структур і закономірностей. Охарактеризовано особливості «глибокого навчання» та фактори його популярності. Виділено каузальні мережі як клас моделей, які поєднують у собі переваги генеративних, ціле-визначених та багатоцільових моделей і відрізняються тим, що придатні для прогнозу ефектів керування (втручання). Вказано шість «опор», на яких будується методологічне ядро великої аналітики.

Ключові слова: великі дані, аналіз даних, виведення моделі, відкриття знань, статистичні методи, предиктивні та генеративні моделі, каузальна мережа, прогноз.

Великі дані. Роль і значення

Великі дані (Big Data) стали помітним феноменом розвитку інформаційних технологій пост-індустріального суспільства і впливають на різні аспекти життєдіяльності, від політики до наукових досліджень [1–12]. Останні 10–15 років позначені безпрецедентним зростанням електронних (комп'ютеризованих) зібрань даних з різноманітних сфер діяльності. Це явище характеризують як повінь даних. Є багато варіантів визначення великих даних; не повторюючи їх, нагадаємо головні особливості, що спонукали ввести поняття великих даних. Великі дані (ВД), по-перше, відрізняються такими величезними обсягами, що їх зберігання, супроводження, управління і доступ до них наштовхується на обмеження існуючих технічних й програмних засобів [13]. Проблеми й перешкоди подекуди мають не тільки технічний, а й принциповий характер. Тож, робота з великими даними потребує нових нетрадиційних рішень. Обсяги накопичених сьогодні електронних даних вимірюються зетабайтами (тобто величинами порядку 10^{21} байт) [4–10]. Деякі дані генеруються (породжуються) з такою швидкістю, що маємо тільки дві альтернативи – або втрачати ці дані, або записувати їх негайно й такими, якими вони виміряні (сприйняті). Тому великі дані часто характеризуються як «швидкі», «сирі», неструктуровані й неточні. Можна сказати, що ВД виникли внаслідок впровадження автоматичних засобів та механізмів, які швидко й майже безупинно вимірюють та реєструють цифрові (електронні) дані з відповідного середовища (обладнання). Результати вимірів автоматично записуються «за годиться», хоча переважна маса даних залишиться не спожитою і згодом буде стерта. Назвемо типові середовища й джерела, звідки походять великі дані. На сьогодні відомі наступні джерела ВД: сенсорні мережі, прилади промислових об'єктів та технологічних ліній виробництва, торгові центри (супермаркети), інфраструктурні системи (енергетичні, транспортні тощо), соціальні мережі в Інтернеті, YouTube з «океаном» відеофайлів, системи on-line продаж, мобільний зв'язок, біржі та інші фінансові центри, навколоземні супутники спостережень, різноманітні датчики та прилади контролю навколишнього середовища та екологічних служб, прилади відео-спостереження, файли зображень з автоматизованих телескопів, устаткування експериментальних досліджень з фізики частинок, прилади біомедичних обстежень (зокрема, МРТ-зобра-

ження), дані біохімічних вимірів (генетика, протеоміка) і т. д. Побіч того, великі дані породжуються в результаті переведення у цифрову форму даних державних, адміністративних та суспільних реєстрів, медичних карток, статистичної звітності й т. д.

Актуальність аналітики великих даних визначається прискоренням збору і накопиченням великих масивів емпіричних даних з різноманітних сфер діяльності суспільства, а також готовністю наукових, програмних й комп'ютерних ресурсів для створення аналітичних продуктів. У розвинутих країнах світу в цих дослідженнях й розробках задіяні величезні ресурси й численні наукові та інженерні кадри [5–11]. Комплекс досліджень та розробок під назвою «великі дані плюс велика аналітика» не є абсолютно новим явищем. Його можна сприймати як продовження (або новий етап) того алгомеративного й інтегративного напрямку розвитку методів, засобів й технологій, який називали Data Mining, Knowledge Discovery in Data, інтелектуальний аналіз даних, виділення знань з даних і т. п. Багато положень, методів та напрацювань, отриманих «під дахом» названих понять, залишаються адекватними й корисними для ВД [14–18]. Водночас існує низка особливостей, що характеризують новизну ВД й великої аналітики. Якщо два десятиліття тому доступ та підготовку даних розглядали як допоміжний етап, то тепер пошук, доставка й попередня обробка великих даних стають все більш проблемним етапом усього циклу використання даних. Пріоритет зусиль зміщується на інструменти й технології пошуку, доступу до потрібних «сирих» даних та підготовки релевантних даних (маніпуляції з ВД).

Діапазон впроваджень великої аналітики охоплює бізнес, державне управління та наукові дослідження. Підтверджується теза про зсув у методології досліджень. Підвищується роль індуктивно-емпіричного підходу. Можна казати, що формується парадигма прискореного пізнання на основі узагальнення емпіричних даних [6, 14, 15, 19–22]. Доступність всебічних релевантних даних дозволяє автоматизувати процес наукового відкриття, і деякі автори проголосили настання четвертої ери в історії науки [21]. Традиційне публічне наукове обговорення залишається необхідним, але його акцент зміщується з етапу висунення гіпотез і процесу вироблення положень і рішень на етап оцінки й інтерпретації результатів (теорії) та їх інтеграції в систему наукових знань.

Витрати на збір і зберігання даних величезного обсягу виправдані тільки якщо ті дані будуть результативно використані і забезпечать достатнє відшкодування (зиск). Деякі потреби можна задовольнити окремими записами, вилученими з великих даних. Мається на увазі, що кінцевим результатом стають відповідні файли, фрагменти чи записи, відібрані з масиву даних в тій формі, в якій вони зберігаються. Наприклад, для розслідування злочину потрібні окремі записи в журналах або кадри відео-спостереження. Проте іноді знайти у ВД потрібну інформацію стандартними засобами (зокрема, через SQL-запити) важко. Справа не тільки у обсягах вмістилища даних і проблемах доступу до них. Часто неможливо точно описати, що саме аналітик (користувач) хоче знайти, важко сформулювати запит. Але головний напрямок використання ВД – іншого характеру. Основний шлях результативного використання ВД здійснюється через глибокий аналіз даних, коли величезний масив сирої інформації перетворюється («перетравлюється»), і на виході видається компактна, концентрована й цінна інформація кінцевого споживання. З даних вилучається (екстрагується) їх цінний сенс. Отже, великі дані «автоматично» передбачають велику аналітику (ВеАн).

Часто організація (фірма, орган управління) має у своєму розпорядженні великі зібрання даних, але ці дані дуже вибірково та обмежено залучаються до процесу досліджень, підготовки планів чи прогнозування наслідків пропонованих управлінських рішень. Вибір і обґрунтування рішень традиційно робилися на основі експертних суджень і оцінок, адекватність й актуальність яких важко контролювати. (Крім того, часто експертні міркування та суб'єктивні уявлення покладають

в основу побудови математичних моделей). Зазвичай адекватна модель є невідома, а знання про об'єкт існують як сукупність розрізнених відомостей та уявлень вузьких спеціалістів. Таку «скирту інформації» важко узгодити та звести у робочу модель.

Доступність ВД дозволяє отримати широкий спектр інформації про об'єкт та середовище. З'являється можливість побудувати замкнений комп'ютеризований цикл планування і керування. Вихід ВД на ринок ІТ-продуктів дозволяє кардинально оновити технологію й практику підготовки й обґрунтування важливих рішень. Нова технологія рішень будується як комп'ютерна, з визначальною роллю даних (data-driven), що дозволяє позбутися консерватизму й суб'єктивізму у керуванні. Рішення для керівництва фірми виробляються як прямий результат аналізу й переробки комплексу різноманітних релевантних даних (наприклад, про процеси продаж, поведінку споживачів, про діяльність підрозділів фірми тощо).

Підготовка планів, прогнозування наслідків рішень і дій, а також інші аналітичні дослідження мають безпосередньо ґрунтуватися на аналізі масивів емпіричних даних. Актуальна задача – ідентифікація потрібної адекватної моделі «об'єктивними» методами на основі зібраних даних спостережень. Шукана модель приречена бути емпіричною (за витоками) та феноменологічною (за змістом і формою подання).

Комп'ютеризація цілого циклу менеджменту має вирішальне значення для маркетингу популярних й стрімко обновлюваних продуктів (гаджетів, засобів побутового комфорту тощо). Сотні тисяч компаній скористалися ресурсами, сервісами та аналітичними засобами Amazon Web Services, побудованими на хмарних технологіях зберігання даних та обчислень. Перелічимо деякі приклади застосувань ВД, згадані в статтях [5–10, 12, 23–26].

В корпорації Шеврон проаналізували терабайти сейсмічних даних мексиканської затоки, поліпшили свої комп'ютерні моделі, і в результаті підвищили успішність буріння від 1 з 5-ти спроб до 1 з 3-х. (Одне буріння вартує 100 мільйонів доларів). Деякі страхові компанії тепер не тільки відстежують добробут й майно клієнтів, а й також збирають дані сенсорів, вмонтованих у автомобілі, й аналізують кілометраж, маршрути, час поїздок тощо. Транспортна компанія U.S. Xpress підтримує моніторинг сенсорних даних про стан та місцезнаходження своїх авто й вантажівок, а також дані з мобільників та гаджетів водіїв та операторів. Дані накопичуються у хмарі й аналізуються. За результатами аналізу оптимізується керування усім автопарком. Вантажівки вчасно спрямовуються до ближчих заправок пального з нижчою ціною. Для техобслуговування авто-засоби спрямовуються до оптимально підібраних депо. Враховуються затори на дорогах, потреба розігріву мотору взимку і т. д. Кілька інших фірм й агентств аналізують великі дані для оптимізації логістики та постачання енергії.

Служби й агентства з охорони здоров'я інтегрують дані з різних джерел. Медичні зводи доповнюються й уточнюються даними індивідуального рівня з соціальних мереж, даними викликів медслужб по мобільним телефонам тощо. Інтеграція клінічних даних з даними поведінки та суспільними показниками допомагає знизити вартість та підвищити якість лікування. Для оцінки заходів з охорони здоров'я у провінції Квебек дослідники зіставили медичні записи з даними продаж продуктів харчування в тому ж регіоні (спираючись на поштовий код). Покупки з використанням карток лояльності дозволили прив'язати споживачів до місць проживання. Зіставлення адміністративних даних з індивідуальними даними (релевантними до стану здоров'я) дозволило уточнити статистику захворювань діабетом. Аналіз соціальних мереж та пошукових слів в Інтернеті дозволив значно підвищити оперативність моніторингу розповсюдження легеневих інфекцій. Відомий факт, що аналіз «твітів» допоміг простежити розповсюдження холери.

Провайдери мобільного сервісу аналізують демографічні дані споживачів,

статус їх житла, деталі користування сервісами, що дозволяє надавати оптимальні персоналізовані пропозиції телекомунікаційного сервісу. Також провайдери розробляють систему оперативного виявлення обману (використовують предиктивну аналітику). Менеджери центрів роздрібної торгівлі тепер аналізують не тільки кошики покупок, але й потоки покупців з розбивкою на соціальні групи. Компанії роздрібної торгівлі on-line (зокрема, Amazon) для персоналізованих рекомендацій аналізують пошукові слова користувача, його «кліки» протягом сеансу, покупки у минулому тощо. Великий список посилань на застосування великих даних можна знайти в [1, 5].

Процес організації великої аналітики (обрис)

Цінність великих «сирих» даних визначається нашою здатністю вилучати з них «сенс», корисний за змістом і зручний за формою. Практика вимагає виділяти цінний екстракт швидко, використовуючи «свіжі» дані. Коли сукупність доступних даних охоплює екстремальне широкий спектр інформації, фірма (організація) може виконувати багато оперативних функцій автоматизовано, майже повністю на основі ВД. Отже, треба будувати замкнений комп'ютеризований цикл технологій – від збору даних до кінцевого застосування результатів (рішень, керування). «Непрозорі» й не-комп'ютерні процедури виносяться за межі «оперативного» циклу керування. (За штабами фірми залишаються функції нагляду (супервізія) та вищий рівень керування.) Виконання аналітичного завдання завершується видачею моделі або результату в формі, придатній для кінцевого застосування. (Вживають термін «actionable outputs».) Такий результат може використовуватися протягом певного періоду, коли виконується «короткий» цикл аналітики (для керування використовують «свіжі» дані звуженої номенклатури). Схема циклів життя ВД та ВеАн (великий цикл, цикл аналізу, цикл використання) зображена на рис. 1.

Оскільки ВеАн використовує переважно статистичні методи, дані мають складатися з списку випадків (прикладів), що характеризують однотипні об'єкти або той самий об'єкт у варіабельних умовах. Випадки можуть трактуватися як екземпляри популяції, прецеденти, транзакції, цикли та періоди функціонування. (Існують дані, де поняття випадків та прикладів не збігаються [16]). Більшість традиційних методів аналізу потребують, щоб дані всіх випадків складалися з єдиного набору ат-

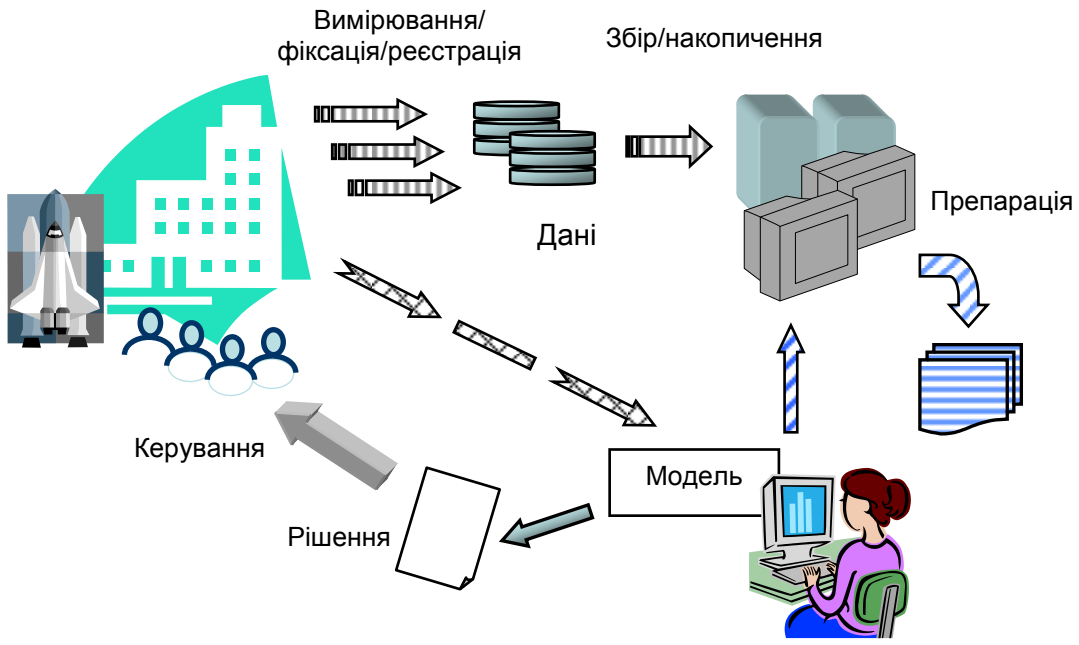

Рис. 1. Цикли великої аналітики і використання даних

рибутів і збиралися за єдиною схемою вимірювання. Більшість класичних методів й процедур аналізу даних розраховані на зручно форматовані дані (зазвичай – у формі таблиці), що вміщуються в пам'яті комп'ютера. Натомість ВД наповнені переважно «сирими», різнорідними, неузгодженими, невпорядкованими та неструктурованими даними. Інформація щодо певного випадку може знаходитися у різних файлах і сховищах. Іноді доводиться розглядати як «випадок» не тільки вектор чисел, а й цілий образ, текст, структуру і т. д. В деяких даних неясно, як розрізнити і виділити окремі випадки.

Дані, що зберігаються у сховищах, можна поділити на: 1) структуровані; 2) «гнучко-структуровані» або слабко-структуровані; 3) неструктуровані. До структурованих відносять дані, організовані за жорсткою схемою. Кожна одиниця (запис) даних складається з уніфікованого набору позицій, і кожну позицію займає елемент (атрибут) відповідного відомого змісту. (Часто це елементи одного типу, наприклад, дійсні числа.) Така структуризація гарантує прості й «прозорі» процедури імпорту даних в усіх платформах. Гнучко-структурованими можна назвати дані, де не зафіксовано набору позицій для елементів. До цього виду належать дані широкого спектру, включаючи довільні послідовності символів, графові структури, мовні тексти й гіпертексти. До гнучко-структурованих треба віднести також дані, які побудовані за рекурсивними схемами (з невизначеними розмірами). Текст має свою структуру, визначену синтаксисом, граматикою та іншими обмеженнями, але така структуризація не забезпечує однозначної інтерпретації елементів (слів) і не підтримується стандартними процедурами обробки. Неструктуровані дані не мають чітко визначеної структури. Для використання неструктурованих даних потрібні нестандартні процедури конверсії, спеціальна розмітка, додаткові дескриптори і т. п. Схожі проблеми виникають, коли дані структуровані, але структура фіксації даних нерегулярна і невідповідна (або невідома аналітику). Маємо проблеми, коли не тільки фізична, але й логічна структура даних не збігається із змістовною («семантичною») структурою. Такі дані виникають, наприклад, коли записується потік сигналів або коли об'ємне зображення описується простою послідовністю точок (пікселів). Можна виділити також дані з частково-невідомою структурою. Деякі дані можна інтерпретувати та «зрозуміти» тільки з допомогою «автора» даних. Неструктурованість та різнорідність даних створює певні проблеми для обробки [6–10, 16, 27, 28]. Потрібні попередні етапи компіляції та інтеграції даних.

Доволі частою є ситуація, коли окремі прилади (засоби) автономно збирають дані про ті самі індивіди популяції чи про ті самі (або еквівалентні) транзакції, цикли функціонування об'єкту, а зібрані дані накопичуються в окремих файлах. Для того, щоб могли працювати типові методи аналізу даних, потрібно співвіднести (ідентифікувати) відповідні записи в різних файлах і сформувати «випадки». Але це не вдається зробити, якщо в файлах немає інформації, яка допомогла би однозначно розпізнати і ототожнити прецеденти (випадки). З точки зору багатовимірного аналізу, маємо ситуацію вертикально-секціонованих, («розщеплених») даних. Отже, з метою отримання з ВД корисного «сенсу» перед власне результативним аналізом необхідно виконати відбір та підготовку даних [7].

Процес великої аналітики включає два етапи:

1) доставка та компіляція даних (пошук, добір, фільтрація, агрегація, комплектування, інтеграція, зменшення розмірності, синхронізація, переформатування);

2) власне глибокий аналіз підготовлених даних.

Ланцюг проходження завдання BeAн показано на рис. 2. Етап глибокого аналізу даних у свою чергу може складатися з ланцюга завдань. Попередня обробка може залучати методи, які традиційно розглядалися як методи власне аналізу даних (аналіз головних компонент, random projection і т. д.).

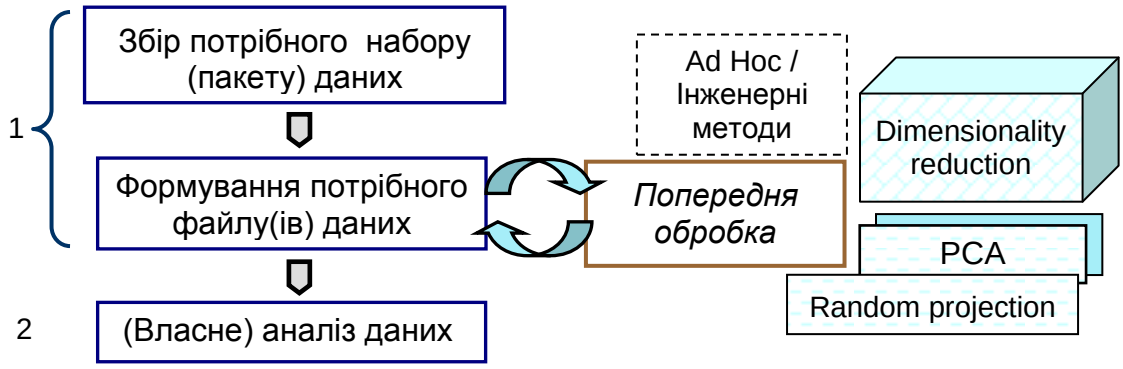

Рис. 2. Схема процесу великої аналітики

Одна з тенденцій BeAн – перенесення аналітичних засобів в програмне забезпечення баз даних, аби виконувати значну частину роботи в місцях зберігання, без передачі даних на сервер аналітика. Зокрема, фірма SAS у співпраці з Oracle та Teradata інтегрує свою аналітики в програмне забезпечення баз даних. Виконання аналітики прямо на платформі баз даних дає можливість розосередити, розподілити виконання задачі й використати паралелізм. (Такий режим може бути вимушеним у зв'язку з захистом даних.) Але такий режим далеко не завжди прийнятний з огляду на розмаїття методів аналізу, інструментарію й мов програмування. До того ж, масовану ітеративну переробку (з багаторазовим скануванням активної порції даних) зазвичай ефективніше виконувати, маючи ректифікований файл в локальній пам'яті комп'ютера аналітика. В деяких платформах та інструментах застосовується режим in-Memory Analytics, коли «гарячі» дані утримуються в пам'яті RAM (не переміщуються на диск).

Існує інструментарій, який інтегрує бази даних підрозділів (департаментів) у єдину систему даних фірми (організації) [16, 17, 26]. Одна з передових сучасних програмних платформ аналізу ВД (яка підтримує увесь цикл аналізу) – Apache Hadoop and MapReduce. Багатий комплект методів і програм аналізу даних для вказаної платформи надається відкритим середовищем Apache Mahout та Apache Spark [7–10, 26, 29]. Популярне відкрите середовище з мовою **R** (наступниця мови **S**) інтенсивно поповнюється програмами розв'язання різноманітних задач. В [30] наводиться низка пакетів програм та застосувань аналітики у біомедичній галузі. В роботі [5] надано перелік фірм-постачальників платформ для ВД.

Перспективні режими та інтелектуалізація аналізу даних

Великі дані надають ресурси і створюють передумови для виходу за межі можливостей стандартних методів пошуку інформації (SQL-запити, ключові слова). На черзі розробка технологій «інтелектуального» пошуку інформації (ІнтлПІ). Результат виконання ІнтлПІ може виглядати традиційно, тобто як запис, фрагмент чи цілий файл даних в тій формі, як він зберігається в базі (звісно, видається композиція знайдених фрагментів даних). Але суттєва відмінність полягає в тому, що користувач може скористатися інтелектуальним пошуком, коли не знає, як сформулювати запит (хоча він знає, що йому потрібне). Йдеться про ситуацію, коли інформаційну потребу не вдається конкретизувати через атрибути та характеристики реальних баз даних. Користувачу недостатньо навіть мета-даних, дескрипторів даних та онтологій для того, щоб специфікувати завдання. Релевантні атрибути та характеристики даних будуть ідентифіковані в процесі виконання ІнтлПІ, тільки після аналізу великого зрізу даних і виявлення певних відношень між багатьма фрагментами даних. Чи є запис даних релевантним залежить не тільки від вмісту цього запису, а й від вмісту інших записів і файлів. Наведемо фік-

тивний приклад гіпотетичного завдання для ІнтлПІ.

«Виявити в якійсь з розвинутих країн сукупність трьох наступних явищ. 1) Невмотивоване різке згортання виробництва сучасного виду озброєння або злам тренду фінансування таких озброєнь (останніми роками). 2) Незадовго перед тим в тій самій країні – суттєві кадрові перестановки в центральному департаменті озброєнь. 3) Одночасно – раптове припинення потоку публікацій з одного з перспективних напрямків науково-технічних досліджень.»

Замість «озброєння» в такому «запиті» могло би фігурувати інше високотехнологічне обладнання. Виконати подібний «запит» автоматично, без участі аналітика, мабуть, неможливо. Спроба звести таке завдання до послідовності SQL-запитів (навіть якби всі дані зберігалися у реляційних базах) була би безперспективною. Намагання відразу формалізувати подібний «запит» напевно призведе до квазі-логічної конструкції з кількома невідомими («вільними» змінними) і нечіткими поняттями. Щоби розпочати виконання запиту, необхідна діалогова взаємодія аналітика з системою. Аналітик має замінити природно-мовний запит пакетом стандартних завдань (де залишаться «вільні» змінні). Як попередню і автономну гілку пошуку можна запустити перегляд преси з метою знайти «відголоски», «сигнали» шуканого явища. («Сигнали» можуть утворюватися як набори таких слів, як «бюджет», «відставка», «полігон», «випробування», «перерозподіл ринку» і т. д.).

Інтелектуальний пошук інформації відрізняється від «інтенсивного» чи «розширеного» режимів традиційного пошуку. В ході «інтенсивного» пошуку аналіз виконується автономно («замкнено») в межах кожного запису чи файлу. (Приклади – пошук особи за фотороботом у великій базі зображень; пошук через Google.) «Розширення» режиму пошуку досягається надбудовою засобів, які враховують розподілену та агломеративну структуру зберігання даних та підтримують прості процедури поповнення даних (наприклад, NoSQL, HDFS та інші) [1, 5, 8, 10, 29]).

Взагалі, можна виділити наступні роди завдань з повномасштабним використанням ВД:

1) розширені режими традиційного пошуку інформації;

2) «інтелектуальний» пошук потрібної інформації (скомпонованих фактів, записів, фрагментів файлів);

3) масована проміжна переробка даних (чи краще сказати – «відпрацювання») однотипною процедурою за один-два проходи через масу даних (mining, concentration, – аналогія із збагаченням руди);

4) індукція моделі об'єкту (джерела), звідки взято дані;

5) екстракція знань з даних (відкриття структур і закономірностей).

«Відпрацювання» даних може бути призначене для підготовки даних перед наступним етапом екстракції знань. Прикладом власне «проміжної» переробки є обчислення достатніх статистик. Альтернативно, якщо на вході задати достатньо інформативні апріорні знання, то в режимі «відпрацювання» можна виробляти кінцевий результат («проміжна» переробка обертається на кінцеву). Далі в огляді в основному розглядаються завдання типу індукції моделей та екстракції знань.

Для проведення повномасштабних емпірико-індуктивних досліджень, які відштовхувалися б від «сирих» даних і доводили результати до рівня «кристалізованих» знань, необхідно побудувати багаторівневу високоорганізовану інтегровану технологію, з адекватними мовами спілкування між рівнями. Така технологія демонструватиме властивості, які вважаються інтелектуальними [20], і зможе підміняти (а в чомусь й перевершувати) людину-аналітика.

Повномасштабний процес виділення знань з даних дозволяє в одному великому (можливо – ітеративному) циклі аналізу здійснити те, що раніше (за посередництва аналітика) виконувалося набором завдань розвідкового (експлоративного) та конфірмативного аналізу даних.

Основні напрямки і задачі великої аналітики

Останніми роками на ринок засобів підтримки бізнесу виходять IT-продукти та інструментарій, що кардинально оновлюють технологію менеджменту та вироблення рішень, ґрунтуючи їх на аналізі ВД. Зрозуміло, найбільше враження справляють такі інструменти й технології, де вихідним результатом аналізу даних є практичний висновок, кінцева рекомендація чи навіть варіант бізнес-рішення. Образно кажучи, найбільш привабливою є переробка даних за схемою «стимули–рефлекс» (подібно до того, як регулюється поведінка тварини або елементарні акти поведінки людини). Тобто на виході технології отримуємо вказівку до дії («actionable output»). Така схема (суцільна «чорна скриня») працює для спеціальних задач, наприклад, розпізнавання (де ціль вказана, а ролі змінних в принципі відомі). Але стосовно проблем управління такий рівень «самостійності» та «самодостатності» інформаційної технологій може бути практично ефективним лише для дуже елементарного рівня «миттєвого» управління або за ідеалізованих умов (наприклад, за надзвичайно високої спеціалізації діяльності фірми у дуже стабільних умовах ринку). Для більшості практичних ситуацій така схема роботи нереалістична. Було б контрпродуктивно намагатися занурити у «чорну скриню» увесь процес вироблення рішень організації, цикл керування складним об'єктом або ціле дослідження. Більш реалістичне й корисне завдання для комп'ютера – знайти в даних цікаву (для користувача) інформацію, виділити закономірності, відкрити знання, побудувати «портрет» об'єкту у середовищі, вивести модель, яка відтворює систему зв'язків та впливів (показує, «як все розгортається»). А вже на основі отриманих «знахідок» (знань) аналітик і користувач зможуть виробляти вказівки до дії.

Задача глибокого аналізу даних визначається метою аналітика (користувача) і типом потрібного результату. Це має вказати аналітик (програміст) [14, 15, 20, 22, 27]. Іноді строго сформулювати завдання і дати постановку задачі важко. (Навіть для таких конкретних задач, як розпізнавання об'єктів чи образів, часто не формулюють строгої постановки.) Тому в багатьох випадках доречно вести мову не про постановку задачі, а про проблемну ситуацію. Коли аналітик остаточно не визначився з постановкою, можна виконувати розвідкові або стандартні завдання (з «підручного меню»).

Традиційно дані є багатовимірною статистичною вибіркою і подаються у формі плаского масиву. Масив даних має «ширину» та довжину («вишину»). В ширину розташовано набір змінних (атрибутів) X. Мабуть, найбільш «загальне» завдання – стисло описати дані в форматі X. Зрозуміло, що аналітика цікавить не буквальний опис даних $\mathbf{X}$, а опис системи змінних, очищений від гамору і випадкових домішків. Отже, типове завдання (з «підручного меню») – вивести модель даних у формі сумісного розподілення ймовірностей $p(X)$. (Цю задачу часто називають unsupervised learning [31].) Результат такого типу вважається «генеративною моделлю» даних в слабкому сенсі (про генеративну модель в сильному сенсі буде далі). Декларувати таке завдання просто, але коли маємо змінні дійсного типу, і тих змінних багато, і не задано параметричної форми для $p(X)$, тоді незрозуміло, як описувати $p(X)$. Крім того, сама по собі модель в формі $p(X)$, як правило, не цікава. Така модель буде цінною й цікавою, тільки якщо розподілення $p(X)$ демонструє яскраві особливості (часткове виродження, специфічну форму, нестандартні ознаки, неочікувані паттерни), і ці особливості можна компактно описати і змістовно (предметно) інтерпретувати. Коли немає чіткої мети, краще не намагатися знайти повний опис $p(X)$, а дати спрощений опис, характеристику $p(X)$ «в загальних рисах». Наприклад, обчислити моменти, парні коваріації і т. п. (подібні характеристики можна обчислити за одне сканування даних). Така задача – сумарізації даних. Для темпоральних даних корисна інформація – основна частота коливань. Задачі сумарізації даних перетинаються із задачами зниження вимірності даних. Іноді ко-

рисно замість прямого опису даних знайти «прообраз» цих даних, виражений через гіпотетичні змінні Z. (Опис даних $p(X)$ можна відтворити майже без втрат через стандартне перетворення «прообразу» $p(Z)$.) Репрезентація через «прообраз» зацікавить аналітика, якщо гіпотетичні змінні Z взаємонезалежні або їх кількість помітно менша за кількість оригінальних змінних X. Компактну репрезентацію даних може надати класична задача аналізу головних компонент (principal component analysis – PCA) [31]. Нелінійний аналог PCA – виділення принципових кривих, принципових поверхонь тощо. Ще один варіант спрощеного не акцентованого аналізу – шукати тільки інтервали максимальних значень ймовірності $p(X)$. Спеціальним випадком такої задачі можна вважати пошук правил асоціації (у разі дискретних змінних) або узагальнених асоціацій. (Втім, така задача постала як суто прикладна (market basket analysis), без загальної постановки. Виведення правил асоціації можна розглядати як задачу виявлення паттернів.)

Масив даних $\mathbf{X}$ не завжди є статистичною вибіркою за схемою i.i.d. Наявні дані можуть походити з різних популяцій (з різних «моделей»). Тоді важливіше виділити компоненти суміші, а не виводити єдину генеративну модель для суміші. Іноді виділити компоненти можна за допомогою кластеризації (теж завдання з підручного меню). В загальному випадку розділити суміш проблематично (функції щільності ймовірностей компонент можуть значною мірою перетинатися.)

Шанси знайти цікаві, корисні, а головне – практично потрібні результати (моделі, регулярності, паттерни) значно зростають, якщо завдання вдало специфіковане і на вході задано адекватну апріорну інформацію. Традиційні методи аналізу даних були розраховані на вибірки даних малого та середнього розміру. В таких ситуаціях отримати корисний результат можна тільки за умови, що «в загальних рисах» модель задана апріорі. Взагалі, для того, щоб отримати на виході переробки даних змістовний та обґрунтований результат, необхідно подати на вхід сукупність знань та емпіричних даних, таку, що вони «в сумі» утворюють достатньо багату інформацію про об'єкт. Різні співвідношення вказаних двох складових вхідної інформації породжують різні проблемні (когнітивні) ситуації, а відтак – і різні роди задач аналізу даних. Для спрощення задачі традиційно задавали обмеження на вході. Що менше на вході апріорних знань й обмежень, то більшим має бути вміст й обсяг даних.

Додаткова інформативність результату відносно вхідної апріорної інформації («додана вартість» на виході) завжди менше, ніж інформація, що міститься в просканованих та оброблених даних. Часто ця засвоєна інформація становить лише мізерну частку інформації, яка пройшла через процесор, внаслідок того, що дані значною мірою не релевантні для задачі (щодо мети), а також через те, що цінний вміст даних занадто захаращений гамором невідомого характеру. Зростання обсягів даних дозволяє розширити можливості виведення моделей з даних, вивести більш точну й адекватну модель, а також обійтися без жорстких обмежень та важких чи ризикованих припущень на вході.

Мабуть, найпростіший і найпоширеніший спосіб визначити мету й акцент завдання – вказати цільову змінну (характеристику, атрибут) y. Зазвичай змінна y присутня в даних, але іноді її приписує аналітик перед виконанням завдання (робить «розмітку»). В задачах класифікації та розпізнавання цільова змінна y дискретна, а в задачах типу регресії – неперервна. Такі задачі (їх іноді називають supervised learning) можна назвати цілеспрямованими або ціле-визначеними («націленими»). Ціле-визначена задача виводить результат (модель) у формі $p(y|X)$ або $y=\Phi(X)$. (Строго кажучи, оскільки маємо $y\in X$, то треба писати $y=\Phi(Z)$, де $Z\subseteq X\setminus\{y\}$.)

Опис $\Phi(\cdot)$ не обов'язково (не завжди) є аналітичною чи явно вираженою функцією. Це може бути алгоритм, процедура чи просто «чорна скриня». Коли цільова змінна y дискретна, модель вигляду $y=\Phi(X)$ називають «дискримінативною», на противагу «генеративній» моделі $p(X)$.

Результат у формі $p(y|X)$ або $y=\Phi(X)$ також часто називають «предиктивна» (predictive, «передбачувальна») модель. Відповідно, (суто формально) кажуть про предиктивну аналітику. Втім, така модель не обов'язково призначена для «передбачень» майбутніх подій. Зазвичай «предикція» спрямована радше назад у часі (наприклад, при класифікації чи розпізнаванні). Тоді краще сказати не предикція, а відтворення значення змінної y. В науковій літературі предиктивними моделями (та методами) називають такі, які мають здатність до узагальнення, тобто претендують на збереження адекватності поза обробленими даними (на відстані від врахованих прикладів). Іншими словами, до «дійсно предиктивних» моделей ставиться вимога, щоб вони забезпечували адекватну екстраполяцію у просторі прикладів (попри те, що ці моделі виведені із скінченої вибірки даних, яка має вибірковий ухил). В спільноті прагматиків побутує ще більш звужене й вимогливе розуміння предиктивних моделей та «предиктивної аналітики». А саме, характеристика «предиктивна» вживається до методів побудови моделей, наближених до практики бізнесу [24, 32]. В цій спільноті предиктивна аналітика розуміється як така, що забезпечує умовне прогнозування наслідків управління об'єктами, передбачення подій та майбутньої поведінки реальних складних динамічних систем, що розвиваються. До речі, виведення результатів типу «actionable outputs» подекуди називають «прескриптивна» аналітика. А до дескриптивної аналітики відноситься сумарізація даних.

Результат у формі $y=\Phi(X)$ передбачає включення в опис тільки необхідних (значущих) факторів (предикторів, ознак, аргументів, коваріат, регресорів). Отже, коли ставиться ціле-визначена задача, майже завжди мається на увазі, що треба виконати відбір значущих змінних серед заданого набору. Відтак, предиктивні (дискримінативні) моделі схильні до меншої розмірності, ніж «генеративні». Більш того, розв'язання багатьох прикладних задач (класифікації, розпізнавання) передбачає формування ознак, тобто із заданих на вході змінних шляхом комбінації й інтеграції формують нові змінні (підвищеного рівня), які входять в кінцеву «модель» $y=\Phi(X)$.

Останніми роками спостерігається справжній бум досліджень і розробок методів так званого «глибокого навчання». В цих задачах задається не тільки цільова змінна y, але й схема моделі (див. далі). Глибокому навчанню можна протиставити глибокий аналіз даних та відкриття знань. До глибокого аналізу даних відносимо групи задач, які мають на меті:

- відтворити «портрет» об'єкту у середовищі, тобто вивести модель, яка «прозоро» інтерпретується і пояснює функціонування об'єкту;
- відкрити структуру в даних, наприклад, ідентифікувати систему зв'язків та впливів між характеристиками об'єкту у середовищі;
- знайти закономірності поведінки системи (об'єкту) – регулярність, періодичність, інваріанти; знайти аномалії.

Коли завдання спрямоване на відкриття знань, аналітик зазвичай не вказує цільову змінну. Але навіть якщо аналітик задав цільову змінну y, його метою не обов'язково є побудова предиктивної (дискримінативної) моделі $y=\Phi(X)$ чи прогнозування значень змінної y для певних випадків (умов). Метою може бути ідентифікація факторів (причин), які об'єктивно визначають значення y. Задачі відкриття знань та виведення моделей з емпіричних даних (з мінімальними припущеннями на вході щодо майбутнього результату) можна назвати індуктивною переробкою даних. Уявлення про індукцію статистичної моделі як відтворення адекватного опису процесу генерації даних започаткував ще Р. Фішер. Адекватний опис процесу генерації даних – це «портрет» джерела даних (а відтак – і об'єкту). Він допомагає зрозуміти, що саме й чому відбувається.

Відмінність між виведенням моделі об'єкту та відкриттям знань в даних можна характеризувати наступним чином. По-перше, задача відкриття знань ставиться з суттєво меншою вагою апріорних знань на вході (принаймні щодо предмету відкрит-

тя). По-друге, модель, як правило, претендує на опис всіх оброблених даних (із застереженням, що коли йдеться про предиктивну модель, вона включає не всі представлені, а тільки релевантні змінні.) Натомість «знання» (результат процесу екстракції знань) є паттернами, які не завжди підтримуються всіма даними, але повторюються достатньо регулярно. (З точки зору статистики, це такі паттерни, що частота їх підтвердження в даних суттєво перевищує рівень, який можна було би пояснити випадковістю.) Закономірність може стосуватися лише окремого зрізу (чи сегменту) даних, проте виконується систематично. Третя відмінність «знань» – вони мають пізнавальну імпресивність. Знання й паттерни відображають яскраві особливості, які легко інтерпретуються і є цікавими в пізнавальному сенсі. Натомість модель у формі $y = \Phi(X)$ може не показувати нічого цікавого, але вона продуктивно «працює».

В результаті глибокого аналізу даних можуть бути знайдені імплікативні правила вигляду: (вектор характеристик $\mathbf{A}$) => (окрема_характеристика_B). Такий результат теж вкладається у форму $b = \Phi(A)$, хоча може залучати не аналітичні, а логічні вирази. Такий результат вважається знахідкою й знанням тільки в тому разі, якщо правило дає значення b з високою точністю (детерміновано), і до того ж змінні b та A були автоматично знайдені, а не задані на вході. Аналогічну ситуацію маємо щодо кластеризації: про відкриття знань доречно говорити тільки якщо знайдені кластери є статистично значущі. Базовий принцип виділення регулярностей – знайдення часто повторюваних сполучень, паттернів, схем, або навпаки, занадто рідких сполучень («дірок» у розподіленні). Критерієм регулярної повторюваності (або регулярної відсутності) є значне відхилення від статистично очікуваних значень (або від очікувань аналітика). Аби отримані результати можна було прийняти як адекватні знання (закономірності) про об'єкт, необхідно запобігти сценарію, коли ті результати (неявно) закладено в процедурах виведення, або коли вони є артефактами збору чи попередньої обробки даних.

На рис. 3 запропоновано один з варіантів систематизації великої аналітики за родами задач та типами результатів.

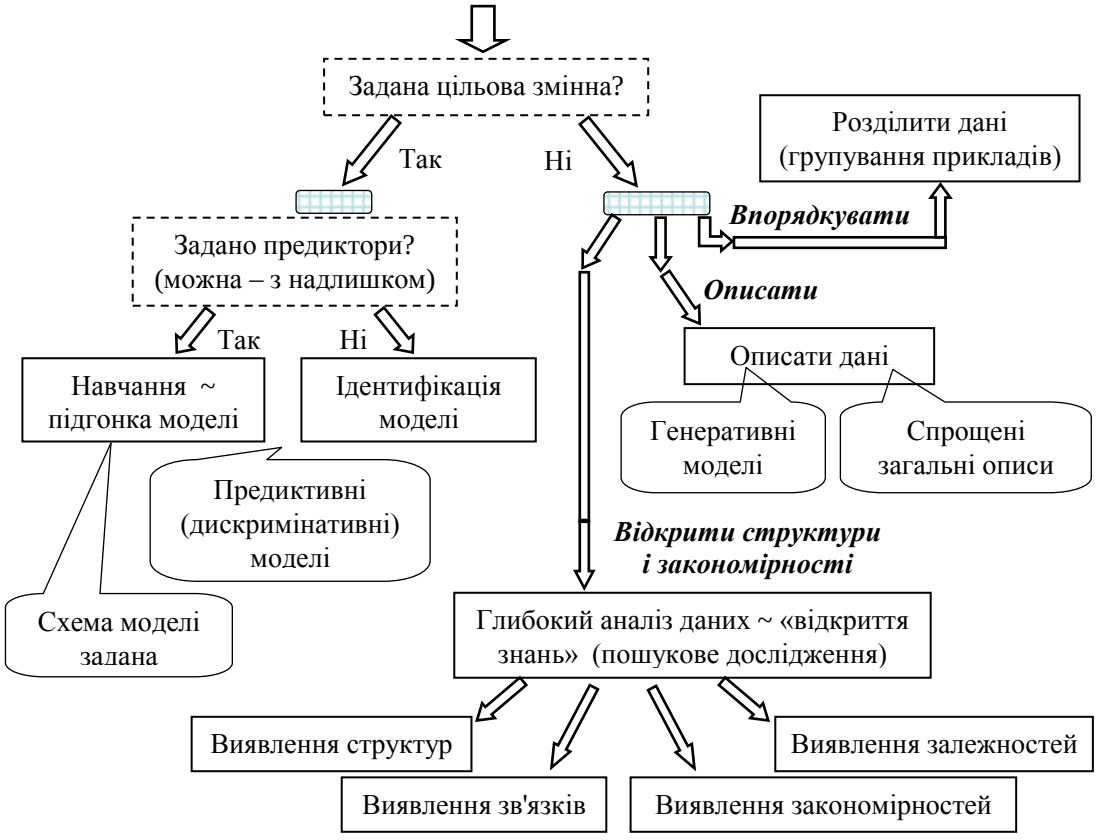

Рис. 3. Типи задач та результатів великої аналітики

Типові задачі аналізу даних

Вибірково окреслимо підходи до аналізу слабко-структурованих даних. Методи аналізу текстів екстрагують потрібну інформацію з новин, оглядів, електронних листів, «твітів», документів, статей. Текст – це не емпіричні дані; текст не задовольняє припущень, прийнятих в традиційних методах аналізу даних. (Втім, великі зібрання документів можуть розглядатися як вибірка екземплярів популяції й аналізуватися статистично.) Залишимо поза увагою комп'ютерну лінгвістику та технології, спеціалізовані на мовах. Прості методи кількісного аналізу тексту виловлюють лише «поверхову» інформацію. Внутрішня структурованість тексту незручна для традиційних методів аналізу. Відомо дві групи простих методів аналізу текстів: 1) екстракція інформації; 2) сумарізація текстів [7]. Методи першої групи розпізнають у тексті об'єкти (сутності) та виділяють відношення між ними. Методи сумарізації текстів застосовують дві техніки. «Екстрактивна сумарізація» робить компіляцію вирізок (фрагментів) заданого тесту, враховуючи місце та частоту входження слів. «Абстрактивна сумарізація» намагається виявити семантику тексту і може видати результат в інших термінах і конструкціях. Щоб автоматично з'ясувати зміст і сенс тексту, залучаються методи обробки природної мови. Розробляються також методи генерування природомовних відповідей на запитання, а також методи розпізнавання опіній та настроїв, які приховані «між рядками» тесту. Методи аналізу даних з Web-середовища можна знайти у [33]. Введення в аналіз даних соціальних медіа дається в [34]. В роботі [35] описано, зокрема, аналіз даних в інформаційних мережах, збагачених текстом. Огляд аналізу даних з Інтернету речей можна знайти в [35, 36].

В аналітичних задачах та у побудові моделей широко вживаються поняття зв'язку і відношення. Але в різних контекстах зв'язок має дуже відмінний сенс. Перелічимо відомі тлумачення поняття зв'язку, які зустрічаються в літературі з комп'ютерних наук та інформаційних технологій. Отже, типи зв'язку: логічна залежність (зчеплення окремих значень); статистична залежність (зчеплення частот значень); суміжність; близькість; посилання («лінки», адресація); (безпосереднє) слідування у часі; відношення «об'єкт – атрибут (ознаки)»; «ціле – частина (деталь)»; відношення приналежності (до класу). Залежність має семантичні градації: асоціація, вплив, каузальний зв'язок.

Можна запропонувати наступний перелік типових задач ВеАн:

1) групування випадків (записів, об'єктів); кластеризація;

2) виведення ціле-визначених моделей (для класифікації, регресії, розпізнавання);

3) виявлення регулярних паттернів (систематичних повторювань):

– структурних, зокрема, послідовних (motifs), 3-вимірних, графових,..

– наборів (правил асоціацій, item sets, market baskets), .. ;

4) виявлення типових (для популяції) дискретних послідовностей у часі (лінки, ланцюги дій тощо);

5) виявлення трендів, періодичності та аномалій (в даних із темпоральною прив'язкою);

6) відтворення структур залежностей;

7) відтворення каузальних моделей.

Впорядковані у часі дані (ряди даних) не є статистичною i.i.d.-вибіркою у буквальному розумінні (хоча за певної трансформації теж можуть розглядатися як стандартна вибірка). Темпоральні дані (в першу чергу для неперервних процесів) надають простір для специфічних задач аналізу, наприклад, виявлення періодичності, трендів, динамічних аномалій [37]. Знайдені тренди та періодичність у даних допомагають виконувати «феноменологічний» безумовний (інерційний) прогноз. Інші регулярні паттерни також допомагають прогнозувати у відповідних ситуаціях. Натомість знання каузальної моделі дає аналітичний інструмент для прогнозування наслідків втручання в об'єкт (керування). Моделі розпізнавання або класи-

фікації (в першу чергу ті, що побудовані як нейронні мережі) радше надають не знання, а вміння. До виявлення знань можна зарахувати хіба що підбір підмножини значущих предикторів. Але більшість традиційних методів розв'язують цю задачу у дуже спрощеному й спеціальному варіанті.

Вузька спеціалізація традиційних ціле-визначених моделей випливає не тільки з фіксації цільової змінної, але й з прив'язки до формату кандидатів у предиктори (фактори). Нехай для заданих y та X була виведена модель $\hat{y} = \Phi(Z)$, де $Z \subseteq X$. Можлива ситуація, коли потрібно оцінити (спрогнозувати) значення y за умови, що відомі значення тільки деяких факторів, тобто змінних Q, причому $Q \subset Z$. Як застосувати модель в цій ситуації? Якщо $\Phi(\cdot)$ – формула, що підставити в формулу на місце невідомих факторів? Якщо $\Phi(\cdot)$ – процедура, що подати на її відповідні входи? Простої задовільної відповіді на ці питання немає. Треба враховувати кореляцію між факторами, а також їх взаємодію всередині моделі. Поставлене питання знаходить коректну відповідь в апараті каузальних мереж, який дозволяє адаптувати модель до будь-якого формату запиту. (Звісно, вказана проблема неактуальна для тих задач розпізнавання, де вхідні дані характеризуються великою надлишковістю та дублюванням. Наприклад, втрата якихось точок (пікселів) зображення компенсується на етапі вироблення ознак за рахунок сусідніх точок.)

Каузальні моделі

Для забезпечення адаптивності моделі до формату запиту потрібно знати адекватну картину зв'язків між всіма задіяними змінними. (Це потрібно також для ідентифікації справжніх причини для заданого ефекту.) Для задач планування та управління потрібна модель, яка допомагає зрозуміти зв'язки та взаємозалежності між окремими субпроцесами у реальному середовищі об'єкту. Бажано, аби виведена модель була придатна для прогнозу наслідків виконання рішень менеджера (керування). Вказаним вимогам відповідають каузальні моделі і, зокрема, каузальні мережі [38–41]. Факторний аналіз та аналіз незалежних компонент (ICA) знаходять структуру як сукупність незалежних прихованих змінних, які спільно (адитивно) формують значення наявних змінних. Натомість каузальні мережі описують структуру безпосередніх впливів між наявними змінними (зазвичай – в умовах неповної спостережуваності). В процесі виведення каузальної моделі з'ясовується (розпізнається) каузальний характер статистичних зв'язків (кореляцій, асоціацій, залежностей) [39, 42–44]. Стислу характеристику властивостей каузальних мереж можна знайти в [19, 38, 44, 45]. Одним з варіантів КМ є кореляційна мережа для фінансової аналітики [46].

Каузальна мережа (КМ) – це модель залежностей між змінними, яка адекватно відображає структуру спрямованих впливів. КМ описується як пара (G, Θ), де G – граф, що специфікує структуру моделі, Θ – параметри, прив'язані до G, які описують кількісний аспект моделі. В практичних задачах використовують структури без орієнтованих циклів (тобто орграф G – ациклонний). Обмежимося класом моделей з одно-орієнтованими ребрами, тобто на основі ординарних ациклонних орграфів (оАОГ). Множина параметрів оАОГ-моделі складається із сукупності локальних параметрів, заданих для кожної змінної. Зокрема, в мережі, що показана на рис. 4, для змінної Y опис може виглядати як $y = \mathrm{f}(x, z, v) + \varepsilon_Y$. Функція $\mathrm{f}(\cdot)$ може мати будь-яку форму, але зручніше мати справу з лінійною залежністю (що автоматично означає адитивність моделі й індивідуальну прив'язку коефіцієнтів до ребер моделі).

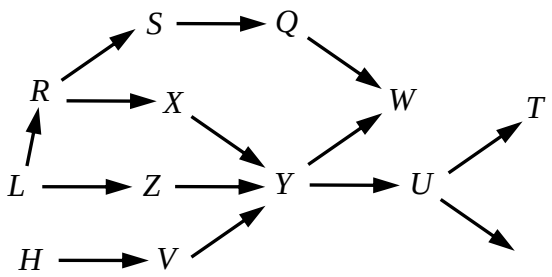

Рис. 4. Приклад каузальної мережі

Каузальні мережі поєднують у собі переваги моделей кількох типів. За умови адекватності, КМ є генеративними моделями в сильному сенсі, – вони адекватно описують процес генерації змінних, ізоморфно («дзеркально») відображаючи процеси в об'єкті. КМ також є предиктивними і дискримінативними моделями, тому що застосовують описи у формі $p(y|X)$ або $y = \mathrm{F}(X)$. Більш того, кожна КМ є багатоцільовою моделлю, оскільки вміщує в собі сукупність ціле-визначених моделей (потенційно – для всіх форматів запиту). КМ можна назвати системою регресійних/класифікаційних моделей, інтегрованих «без швів» за допомогою відношень умовної незалежності.

Каузальні моделі допомагають висвітлити принципову відмінність впливу та асоціації, уточнюють роль та інформативність змінних. Розрізняються два режими використання моделі – «пасивна» предикція та каузальний прогноз («активна предикція»). «Пасивна» предикції розуміється як обчислення значення цільової змінної C, виходячи з значень асоційованих (пов'язаних) з нею змінних $A,B,\ldots$. Така задача формулюється як $p(C|a,b,..)$. Це звичний режим застосування традиційних моделей, зокрема, класифікації. Для пасивної предикції безумовно інформативними виступають всі суміжні змінні, безвідносно до характеру зв'язків (і причини, і наслідки). Наприклад, для моделі, що показана на рис. 4, для предикції (оцінки) значення Y інформативними є X, Z, V, U, W. Але корисний внесок в пасивний прогноз Y можуть зробити й несуміжні змінні, за умови присутності (відсутності) інших змінних у переліку заданих. Умовно- інформативними для Y є всі змінні, поєднані з Y якимось шляхом. Зокрема, якщо не задано значення змінних X, Z, то інформативними стають L та S. Що стосується змінної Q, то вона стає інформативною, якщо задано значення змінної W. (Звісно, внесок «далеких» змінних – незначний). Для пасивного прогнозу корисні не тільки справжні причини, а й тісно пов'язані з ними індикатори. Для класифікації часто використовуються змінні, які радше є наслідками («дітьми») або «братами» цільової змінної.

Каузальний прогноз відповідає на питання, яким буде значення заданої змінної, якщо маніпулювати (керувати) певними іншими змінними (точніше, їх прототипами в об'єкті). Для моделі, показаної на рис. 4, керування змінними Q,W,T не дасть ефекту для Y за жодних умов. Каузальний прогноз значення Y після втручання на змінну X потребує «усунення» внеску конфаудера L. Це здійснюється як корекція моделі (видалення зв'язку $R \to X$). Взагалі, каузальний прогноз для C за втручання на змінну A формулюється як $p(C|do(a),b,..)$ [38–41, 43]. Отже, каузальні мережі є предиктивними моделями у сильному сенсі.

Завважимо, що КМ утворюється зі змінних, заданих на вході, тож і прогноз виражається через них. Але на основі результатів, отриманих з моделі, можна обчислювати «кінцевий» (з точки зору замовника) ефект, для чого залучаються додаткові («зовнішні») функції і фактори, що залишилася поза вхідними даними. (Методи відтворення каузальних мереж з даних згодом будуть розглянуті детальніше.)

Самонавчання алгоритмів та глибоке навчання

У спеціальній літературі часто вживається термін Machine Learning, який зазвичай перекладають буквально – машинне навчання. Під гаслом Machine Learning велися розробки алгоритмів, процедур, методів і програмних засобів розв'язання практичних задач протягом майже пів століття [47]. Ці розробки зосереджувалися на ціле-визначених задачах (оцінка успішності навчання потребує задану ціль). Напрямок Machine Learning (ML) окреслився після того, як дослідники й інженери зрозуміли, що для багатьох прикладних задач (зокрема, класифікації) важко придумати (вибрати) ефективний алгоритм розв'язання. З'ясувалося, що замість того, щоб «вручну» специфікувати потрібний алгоритм розв'язання, краще вирішити задачу вищого рівня – задачу адаптивного конструювання потрібного алгоритму самим

комп'ютером. Тобто запускається автоматичний процес конструювання «цільового» алгоритму як послідовність вибору опцій та параметрів в ході пробних застосувань алгоритму розв'язання кінцевої задачі на «прикладах». Приклади задаються вхідними даними. Підбір опцій диктується успішністю розв'язання прикладної задачі, а весь цей процес називається навчанням. Отже, предметом того, що позначають терміном Machine Learning, є способи і методи автоматичного формування («навчання») алгоритмів і засобів розв'язання прикладних задач на основі досвіду їх розв'язання на прикладах. Коротко це можна назвати «самонавчання алгоритмів» (сНАлг).

Словосполучення Machine Learning широко розповсюдилося в літературі. Щодо вживання «машинне навчання» як терміну можна зауважити наступне. По-перше, воно може дезорієнтувати, бо таке словосполучення стосується також застосування комп'ютерів у навчальному процесі. По-друге, вживання слова «машина» тут не є влучним.

Результатом виконання сНАлг зазвичай є алгоритм обчислення y на основі X (хоча іноді може бути видана модель у певній декларативній формі). Напрямок сНАлг сприймався як такий, що входив до комп'ютерних наук (методів програмування і обчислень) і часто позиціонувався «під дахом» напрямку «штучний інтелект». (До речі, в розвитку самого штучного інтелекту пріоритет змістився від «вилучення» знань (тобто отримання їх від експерта) до виведення знань з даних.) В руслі робіт з сНАлг було винайдено багато способів, тактик, правил, методик й методів, переважно інженерно-евристичних [16, 17, 47–49]. Зокрема, розвинуто інструментарій нейронних мереж. Часто розробники обходилися без математичної постановки задачі, і тільки останнім часом почали запозичувати зі статистики принципи та підходи для обґрунтування, оцінки статистичної значущості та оцінки надійності.

В руслі напрямку нейромереж сформувалася гілка методів так званого «глибокого навчання». Методи глибокого навчання застосовуються переважно для візуального та звукового розпізнавання [50]. В цьому підході опріч цільової змінної y (якою зазвичай є клас об'єкту або «сигнальна» характеристика розпізнавання), за замовчуванням задається й інші апріорна інформація. Вхідні характеристики є кандидатами у предиктори (ознаки) або радше їх компонентами (елементами). Задано також форми перетворення (параметричні родини моделей) або арсенал «цеглин» (будівельних блоків), з яких можна конструювати «модель» $y = \Phi(X)$. Часто задано параметри конструкції «моделі» (кількість рівнів, блоків). Список кандидатів у предиктори (фактори) може бути надлишковим, але зазвичай всі кандидати однакові зі рівнем деталізації і мають однаковий «статус» (це зрозуміло за «фізичним» змістом). Висока спеціалізація «моделі» дозволяє добре специфікувати завдання. По-суті, для глибокого навчання задано «каркас» моделі. «Глибина» в цьому підході означає ієрархічність, багаторівневість конструкції, а також складність використаних функцій (формул). Характер даних диктує необхідність спочатку сформувати з вхідних змінних більш інформативні масштабні ознаки, на основі яких вже побудувати модель. Глибоке навчання продемонструвало, що на відповідному класі задач можна натренувати багаторівневі конструкції, які адекватні при застосуванні до нових прикладів (об'єктів). Успішність глибокого навчання пояснюється характером проблемної ситуації, а саме, наступними обставинами. Вхідні змінні – це дуже «дрібні ознаки» (маленькі частинки «картини», наприклад, пікселі зображень). Велика кількість змінних, причому «сусідні» змінні тісно корельовані і майже ідентичні. Модель високоспеціалізована, з лаконічним результатом на виході (одне з кількох значень). На вході задано «каркас» моделі.

У глибокому навчанні «узагальнення» має сенс об'єднання деталей у ціле, а у глибокому аналізі – радше перехід від одиничного до загального. У глибокому навчанні глибина розуміється як багаторівневість і складність конструкції. Натомість у глибокому аналізі

даних глибина розуміється як сходження від «сирих» випадкових даних до «знань» (до очищеної зрозумілої «картини»), причому ті «знання» випливають з відносин між змінними, характер і роль яких невідомі (змінні можуть бути дуже різнорідними). Ці два напрями різняться також характером переробки даних: перший – це тренування, «підгонка» й оптимізація; другий має пошуково-дослідницький характер [16].

Огляд основних методів аналізу та особливостей їх застосування до великих даних буде презентовано у наступній статті.

Велика аналітика. Проміжні підсумки

Великі дані є одним зі знакових трендів новітніх інформаційних технологій у розвинутих країнах. Великі дані породжуються швидкісними автоматичними засобами реєстрації інформації, вбудованими в реальні об'єкти. Витрати на збір та зберігання великих даних виправдовуються їх результативним використанням, в першу чергу – через глибокий аналіз даних, коли величезний масив сирих даних перетворюється («перетравлюється») на компактну, концентровану й цінну інформацію кінцевого споживання. Аналіз може бути глибоким тільки коли є багата і рясна «сировина».

Взагалі, великі дані можуть бути використані у наступних режимах: «інтелектуальний» пошук інформації; масована переробка даних («відпрацювання», concentration, mining) за один-два проходи; виведення моделі об'єкту (джерела) з даних; екстракція знань з даних (відкриття закономірностей).

Деякі фірми вже впроваджують замкнені комп'ютеризовані технології, що охоплюються увесь цикл оперативного керування – від збору даних до кінцевого застосування (рішень). Великі дані є родючою сировиною для глибокого аналізу (принаймні для аналізу зв'язків) тільки коли вони багатовимірні. Великі дані в принципі можуть забезпечити інформацію, достатню для планування і знайдення оптимальних рішень. Проте потенційна «повнота» даних часто залишається «віртуальною». Великі дані часто є неструктурованими, «гнучко-структурованими» або слабко-структурованими. Крім того, великі дані часто є вертикально-секціонованими («розщепленими»). Перед власне результативним аналізом необхідно виконати підготовку даних. Цей етап може включати такі процедури, як пошук, добір, доставка, фільтрація, агрегація, інтеграція, синхронізація, пере-форматування. Водночас іноді потрібно зменшувати вимірність даних (без втрати їх змістовності).

Можна очікувати, що у майбутньому технології збору даних прогресують, пристрої стануть «тямущими», а інфраструктура розростеться у масштабах. Це забезпечить постачання багатовимірних інтегрованих даних, готових для негайного аналізу. Проте проникнення таких засобів у життя суспільства буде входити у суперечність з правом на приватність й конфіденційність.

Велика аналітика увібрала багатий арсенал кількох дисциплін та набутки різних напрямків розробок. Вона спирається на фундамент статистичної методології (включаючи розвідковий та конфірматив-ний аналіз даних), методи оптимізації та пошуку, методи репрезентації знань та візуалізації багатовимірних даних. Адаптується досвід таких напрямків, як відкриття знань в даних (Data Mining, Knowledge Discovery in Data) і методи самонавчання алгоритмів (Machine Learning). Кілька напрямків досліджень і розробок стали опорами і складовими великої аналітики (рис. 5). Їх об'єднання і взаємне збагачення утворює методологічне ядро великої аналітики.

Типові класи задач аналітики включають: розділення даних (групування випадків); поверховий («загальний») не акцентований опис даних; виведення цілевизначених моделей; відкриття структур та закономірностей. Ціле-визначені задачі охоплюють виведення предиктивних (дискримінативних) моделей, які описують цільову змінну через інші змінні.

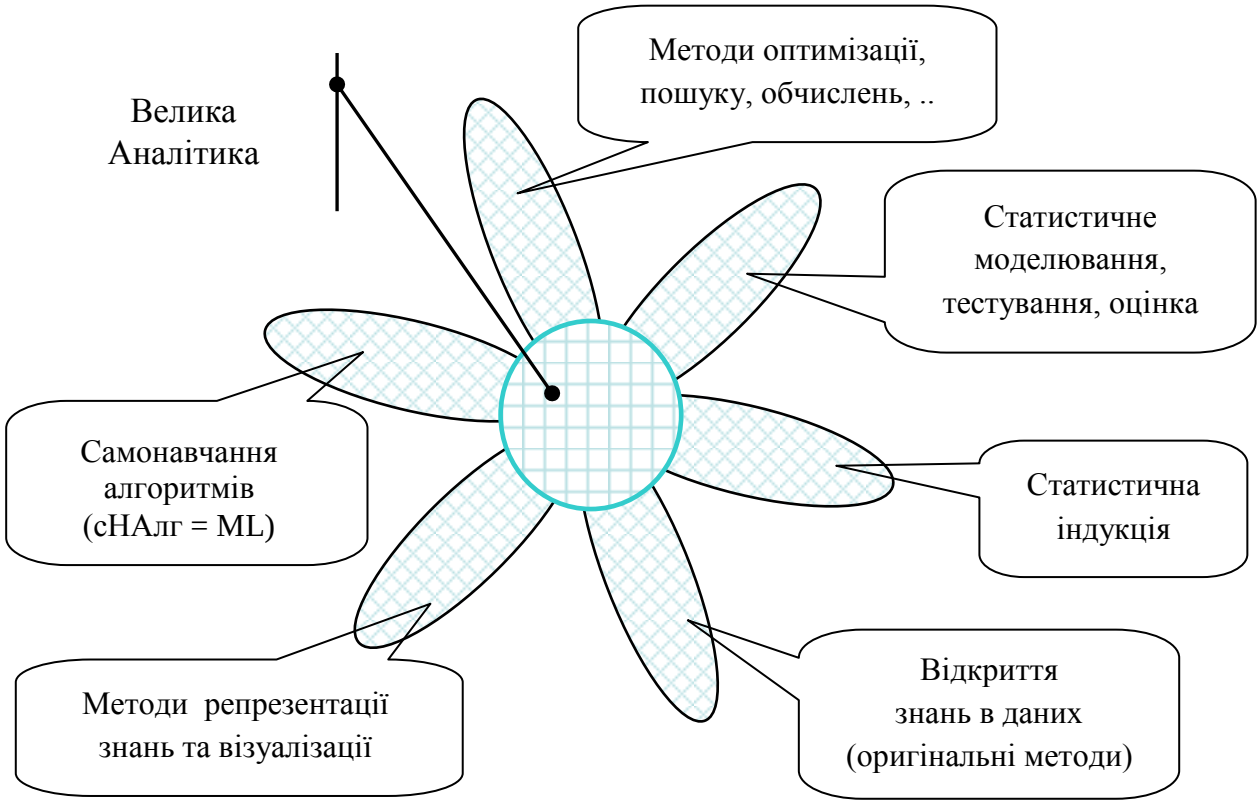

Рис. 5. Фундамент та арсенал великої аналітики

Відмінність моделі об'єкту (результату виведення) і «знання» (результату відкриття) характеризується в трьох аспектах. 1) На вході задачі відкриття знань задається менше апріорної не емпіричної інформації. 2) Модель, як правило, претендує на опис всіх оброблених даних, в той час як «знання» може стосуватися лише окремого зрізу (чи сегменту) даних (проте підтверджується систематично). 3) «Знання» змістовно інтерпретуються, цікаве, неочікуване або надає «інсайт», водночас як модель виконує передбачену функцію або описує дані.

Критичним питанням для адекватності ціле-визначених моделей є підбір значущих предикторів. Модель з високою предиктивною ефективністю не завжди дає розуміння (пояснення) предмету. Популярним різновидом ціле-визначених задач є так зване «глибоке навчання», призначене для розпізнавання образів та мови. Успішність «глибокого навчання» пояснюється спеціальним характером задачі розпізнавання та вхідних даних. Глибокому навчанню можна протиставити глибокий аналіз даних та відкриття знань. У «глибокому навчанні» глибина розуміється як багаторівневість і складність конструкції, а у глибокому аналізі даних – як сходження від «сирих» випадкових даних до «знань», причому ті «знання» не є артефактами алгоритмів виводу чи збору даних, а є результатом «кристалізації» зв'язків, розчинених в масі даних. Форми виявлених закономірностей включають: послідовнісні повторювання (motifs), періодичність коливань індикаторів у часі, інваріанти на основі комбінації характеристик, часто повторювані набори (асоціації), структури залежностей тощо.

Каузальні мережі є генеративними моделями в сильному сенсі, бо вони здатні адекватно описати процес генерації змінних, «дзеркально» відображаючи процеси в об'єкті. Каузальні моделі пристосовані для застосування в режимі варіювання набору заданих значень предикторів (умов). Головна перевага каузальних моделей над традиційними – вони підтримують прогнозування наслідків втручання в об'єкт (керування).

Великі дані надають нові можливості для статистичних методів аналізу і вод-

ночас висувають вимоги до них [16, 23, 24, 28, 30, 31, 32, 51–60]. Результати аналізу великих даних потребують оцінки й верифікації за статистичними принципами. Розповсюдження великих даних стимулює подальший розвиток методів аналізу (зокрема, статистичних) та прогрес комп'ютерних технологій.

Література

1. Big data analytics: a survey. Tsai C.-W., Lai C.-F., Chao H.-C. and Vasilakos A.V. *Journal of Big Data*. 2015. Vol. 2, N. 1. P. 1–32.
2. Science in the petabyte era. *Nature* (journal). 2008. Vol. 455, Issue 7209. Springer Nature Ltd.
3. Frankel F., Reid R. Big data: Distilling meaning from data. *Nature*. Vol. 455, September 2008. p. 30.
4. Doctorow C. Big data: Welcome to the petacentre. Ibid. P. 16–21.
5. Chen C.L.P. and Zhang C.-Y. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*. 2014. Vol. 275. P. 314–347.
6. Cukier K. Data, data everywhere: A special report on managing information. *The Economist*. 2010, February 25.
7. Gandomi A. and Haider M. Beyond the hype: Big data concepts, methods, and analytics. *Intern. Jour. of Information Management*. 2015, Vol. 35, N. 2. P. 137–144.
8. Watson H.J. Tutorial: Big Data analytics: Concepts, technologies, and applications. *Comm. of the Association for Information Systems*. 2014. Vol. 34, Article 65. P. 1247–1268.
9. Sivarajah U., Kamal M.M., Irani Z. and Weerakkody V. Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*. 2017. Vol. 70. P. 263–286.
10. Bhadani A. and Jothimani D. Big Data: Challenges, opportunities and realities / In.: M.K. Singh and D.G. Kumar (eds.). Effective Big Data management and opportunities for implementation. IGI Global, USA, 2016.
11. Intern. Journal of Data Science and Analytics. Special issue on Data Science in Europe. 2018. Vol. 6, Issue 3. P. 163–269.
12. Intern. J. of Data Science and Analytics. Spec. issue on environmental and geospatial data analytics. 2018. Vol. 5, Issue 2–3. P. 81–211.
13. Jacobs A. The pathologies of big data. *Comm. of the ACM*. 2009, Vol. 52, Issue 8, P. 36–44.
14. Андон Ф.И., Балабанов А.С. Выявление знаний и изыскания в базах данных: подходы, модели, методы и системы (обзор). *Проблемы программирования*. 2000, № 1–2. С. 513–526.
15. Балабанов А.С. Выделение знаний из баз данных – передовые компьютерные технологии интеллектуального анализа данных. *Математичні машини і системи*. 2001, № 1–2. С. 40–54.
16. Data mining: practical machine learning tools and techniques / I.H. Witten, F. Eibe, M.A. Hall. (3rd ed.). Morgan Kaufmann, San Francisco, CA. 2011. 629 p.
17. Data Mining. A Knowledge Discovery Approach. K.J. Cios, W. Pedrycz, R.W. Swiniarski and L.A. Kurgan. Springer, 2007, 606 p.
18. Azzalini A. and Scarpa B. Data analysis and Data Mining: An introduction. Oxford University Press, N.Y., 2012. 288 p.
19. Андон Ф.И., Балабанов А.С. Структурные статистические модели: инструмент познания и моделирования. *Системні дослідження та інформаційні технології*. 2007, № 1. С. 79–98.
20. Балабанов О.С. Комп'ютерний інтелект: фантастичні перспективи і щоденний поступ. 1997, revised 2007. [Електронний ресурс.] Доступ: https://www.researchgate.net/publication/332269445_KOMP'UTERNIJ_INTELEKT_FANTASTICNI_PERSPEKTIVI_I_SODENNIJ_POSTUP
21. Hey T, Tansley S. and Tolle K. The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, Redmont, WA. October 2009. 252 p.
22. Siebes A. Data science as a language: challenges for computer science — a position paper. *Intern. J. of Data Science and Analytics*. 2018. Vol. 6. P. 177–187.
23. Fan J., Han F. and Liu H. Challenges of Big Data analysis. *Nat. Scient. Rev*. 2014. Vol. 1, N. 2. P. 293–314.
24. Statistical inference, learning and models in Big Data / B. Franke, J.-F. Plante, R. Roscher, E.A. Lee, C. Smyth, A. Hatefi, F. Chen, E. Gil, A.G. Schwing, A. Selvitella, M.M. Hoffman, R. Grosse, D. Hendricks and N. Reid. *Intern. Statistical Review*. 2016. Vol. 84, N 3. P. 371–389.

25. Swanson N.R. and Xiong W. Big Data analytics in economics: What have we learned so far, and where should we go from here? *Canadian Journal of Economics*. 2018. Vol. 51, Issue 3. P. 695–746.
26. The anatomy of big data computing / R. Kune, P. K. Konugurthi, A. Agarwal, R.R. Chillarige and R. Buyya. *Software: Practice and Experience*. 2016, Vol. 46. P. 79–105.
27. Smirnova E., Ivanescu A., Bai J., Crainiceanu C.M. A practical guide to big data. *Statistics and Probability Letters*. 2018. Vol. 136. P. 25–29.
28. Shi J.Q. How do statisticians analyse big data — our story. *Statistics and Probability Letters*. 2018. Vol. 136. P. 130–133.
29. Jiang H., Chen Y., Qiao Z., Weng T. H. and Li K.C. Scaling up MapReduce-based big data processing on multi-GPU systems. *Cluster Computing*. 2015. Vol. 18, N. 1. P. 369–383.
30. Haughton D. Software packages for data mining. *Wiley StatsRef: Statistics Reference Online*. 2016. P. 1–5.
31. James G., Witten D., Hastie T. and Tibshirani R. An introduction to statistical learning with applications in R. Springer, N.Y., 2013. 426 p.
32. Graham E. and Timmermann A. Forecasting in Economics and Finance. *Annual Review of Economics*. 2016. Vol. 8. P. 81–110.
33. Liu B. Web data mining: Exploring hyperlinks, contents, and usage data. Springer-Verlag: Berlin-Heidelberg, 2011. 622 p.
34. Zafarani R., Abbasi M.A. and Liu H. Social media mining. An introduction. Cambridge University Press. 2019. 380 p.
35. Big Data Analysis: New Algorithms for a New Society. N. Japkowicz and J. Stefanowski (eds.), Springer, Switzerland. 2016. 329 p.
36. Data mining for the Internet of things: Literature review and challenges. F. Chen, P. Deng, J. Wan, D. Zhang. *Intern. Journal of Distributed Sensor Networks*. Vol. 2015. 14 p.
37. Esling P. and Agón C. Time-series data mining. *ACM Computing Surveys*. 2012. Vol. 45, Issue 1. P. 12–34.
38. Pearl J. Causality: models, reasoning, and inference. Cambridge: Cambridge Univ. Press. 2000. 526 p.
39. Spirtes P., Glymour C. and Scheines R. Causation, prediction and search. New York: MIT Press, 2001. 543 p.
40. Балабанов О.С. Відкриття знань в даних та каузальні моделі в аналітичних інформаційних технологіях. *Проблеми програмування*. 2017, № 3. С. 96−112.
41. Peters J., Janzing D. and Schölkopf B. Elements of Causal Inference. Foundations and Learning Algorithms. MIT Press, Cambridge, MA, USA, 2017. 265 p.
42. Shiffrin R.M. Drawing causal inference from Big Data. *Proc. Nat. Acad. Scien. USA*. 2016. Vol. 113, N. 27. P. 7308–7309.
43. Pearl J. and Bareinboim E. External validity: From do-calculus to transportability across populations. *Statistical Science*. 2014. Vol. 29, N 4. P. 579–595.
44. Балабанов О.С. Від коваріацій до каузальності. Відкриття структур залежностей в даних. *Системні дослідження та інформаційні технології*. 2011, № 4. С. 104–118.
45. Балабанов О.С. Відтворення каузальних мереж на основі аналізу марковських властивостей. *Математичні машини та системи*. 2016, № 1. С.16–26.
46. Giudici P. Financial data science. *Statistics and Probability Letters*. 2018. Vol. 136. P. 160–164.
47. Machine learning. Special issue on applications of machine learning and the knowledge discovery process. R. Kohavi, F. Provost. (Eds.) *Machine Learning*. 1998. Vol. 30, N.2/3. P. 127–274.
48. 22nd SIGKDD Conference on Knowledge Discovery and Data Mining, August 13–17, 2016. San Francisco, California.
49. 24th SIGKDD Conference on Knowledge Discovery and Data Mining, August 19–23, 2018. London, UK.
50. LeCun Y., Bengio Y., Hinton G. Deep learning. *Nature*. 2015. Vol. 521. P. 436–444.
51. Donoho D.L. 50 Years of Data Science. *Journal of Computational and Graphical Statistics*. 2017. Vol. 26, Issue 4. P. 745–766.
52. Bühlmann P. and van de Geer S. Statistics for high-dimensional data: Methods, theory and applications. Springer, 2011. 556 p.
53. Bühlmann P. and van de Geer S. Statistics for big data: A perspective. *Statistics and Probability Letters*. 2018. Vol. 136. P. 37–41.
54. Secchi P. On the role of statistics in the era of big data: A call for a debate. Ibid. P. 10–14.
55. Quarteroni A. The role of statistics in the era of big data: A computational scientist' perspective. Ibid. P. 63–67.
56. Cox D.R., Kartsonaki C., Keogh R.H. Big data: Some statistical issues. Ibid. P. 111−115.
57. James G. M. Statistics within business in the era of big data. Ibid. P. 155–159.

58. Weihs C. and Ickstadt K. Data Science: the impact of statistics. *Intern. Journal of Data Science and Analytics*. 2018. Vol. 6. P. 189–194.
59. Efron B. and Hastie T. Computer age statistical inference. Cambridge University Press, N.Y., 2016. 475 p.
60. Carmichael I. and Marron J.S. Data science vs. statistics: two cultures? *Japanese Journal of Statistics and Data Science*. 2018. Vol. 1, Issue 1. P. 117–138.

References

1. Big data analytics: a survey. Tsai C.-W., Lai C.-F., Chao H.-C. and Vasilakos A.V. *Journal of Big Data*. 2015. Vol. 2, N. 1. P. 1–32.
2. Science in the petabyte era. *Nature* (journal). 2008. Vol. 455, Issue 7209. Springer Nature Ltd.
3. Frankel F., Reid R. Big data: Distilling meaning from data. *Nature*. Vol. 455, September 2008. p. 30.
4. Doctorow C. Big data: Welcome to the petacentre. Ibid. P. 16–21.
5. Chen C.L.P. and Zhang C.-Y. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*. 2014. Vol. 275. P. 314–347.
6. Cukier K. Data, data everywhere: A special report on managing information. *The Economist*. 2010, February 25.
7. Gandomi A. and Haider M. Beyond the hype: Big data concepts, methods, and analytics. *Intern. Jour. of Information Management*. 2015, Vol. 35, N. 2. P. 137–144.
8. Watson H.J. Tutorial: Big Data analytics: Concepts, technologies, and applications. *Comm. of the Association for Information Systems*. 2014. Vol. 34, Article 65. P. 1247–1268.
9. Sivarajah U., Kamal M.M., Irani Z. and Weerakkody V. Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*. 2017. Vol. 70. P. 263–286.
10. Bhadani A. and Jothimani D. Big Data: Challenges, opportunities and realities / In.: M.K. Singh and D.G. Kumar (eds.). Effective Big Data management and opportunities for implementation. IGI Global, USA, 2016.
11. Intern. Journal of Data Science and Analytics. Special issue on Data Science in Europe. 2018. Vol. 6, Issue 3. P. 163–269.
12. Intern. J. of Data Science and Analytics. Spec. issue on environmental and geospatial data analytics. 2018. Vol. 5, Issue 2–3. P. 81–211.
13. Jacobs A. The pathologies of big data. *Comm. of the ACM*. 2009, Vol. 52, Issue 8, P. 36–44.
14. Andon P.I. and Balabanov O.S. (2000). Vyjavlenie znanij i izyskanija v bazah dannyh. Podhody, modeli, metody i sistemy. [Knowledge discovery and exploration in databases. Approaches, models, methods and systems]. Problems in programming. N 1–2, P. 513–526. [In Russian]
15. Balabanov O.S. (2001). Knowledge extraction from databases – advanced computer technologies for intellectual data analysis. Mathematical Machines and Systems. N 1–2. P. 40–54. [In Ukrainian]
16. Data mining: practical machine learning tools and techniques / I.H. Witten, F. Eibe, M.A. Hall. (3rd ed.). Morgan Kaufmann, San Francisco, CA. 2011. 629 p.
17. Data Mining. A Knowledge Discovery Approach. K.J. Cios, W. Pedrycz, R.W. Swiniarski and L.A. Kurgan. Springer, 2007, 606 p.
18. Azzalini A. and Scarpa B. Data analysis and Data Mining: An introduction. Oxford University Press, N.Y., 2012. 288 p.
19. Andon P.I. and Balabanov O.S. (2007). Structured statistical models: a tool for cognition and modelling. System Research and Information Technologies. N 1. P. 79–98. [In Russian]
20. Balabanov O.S. (1997). Computer's intelligence: fantastic perspectives and regular progression. Revised 2007. [In Ukrainian] [Electronic resource:] Access: https://www.researchgate.net/publication/332269445_KOMP'UTERNIJ_INTELEKT_FANTASTICNI_PERSPEKTIVI_I_SODENNIJ_POSTUP
21. Hey T, Tansley S. and Tolle K. The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, Redmont, WA. October 2009. 252 p.
22. Siebes A. Data science as a language: challenges for computer science — a position paper. *Intern. J. of Data Science and Analytics*. 2018. Vol. 6. P. 177–187.
23. Fan J., Han F. and Liu H. Challenges of Big Data analysis. *Nat. Scient. Rev*. 2014. Vol. 1, N. 2. P. 293–314.

24. Statistical inference, learning and models in Big Data / B. Franke, J.-F. Plante, R. Roscher, E.A. Lee, C. Smyth, A. Hatefi, F. Chen, E. Gil, A.G. Schwing, A. Selvitella, M.M. Hoffman, R. Grosse, D. Hendricks and N. Reid. *Intern. Statistical Review*. 2016. Vol. 84, N 3. P. 371–389.
25. Swanson N.R. and Xiong W. Big Data analytics in economics: What have we learned so far, and where should we go from here? *Canadian Journal of Economics*. 2018. Vol. 51, Issue 3. P. 695–746.
26. The anatomy of big data computing / R. Kune, P. K. Konugurthi, A. Agarwal, R.R. Chillarige and R. Buyya. *Software: Practice and Experience*. 2016, Vol. 46. P. 79–105.
27. Smirnova E., Ivanescu A., Bai J., Crainiceanu C.M. A practical guide to big data. *Statistics and Probability Letters*. 2018. Vol. 136. P. 25–29.
28. Shi J.Q. How do statisticians analyse big data — our story. *Statistics and Probability Letters*. 2018. Vol. 136. P. 130–133.
29. Jiang H., Chen Y., Qiao Z., Weng T. H. and Li K.C. Scaling up MapReduce-based big data processing on multi-GPU systems. *Cluster Computing*. 2015. Vol. 18, N. 1. P. 369–383.
30. Haughton D. Software packages for data mining. *Wiley StatsRef: Statistics Reference Online*. 2016. P. 1–5.
31. James G., Witten D., Hastie T. and Tibshirani R. An introduction to statistical learning with applications in R. Springer, N.Y., 2013. 426 p.
32. Graham E. and Timmermann A. Forecasting in Economics and Finance. *Annual Review of Economics*. 2016. Vol. 8. P. 81–110.
33. Liu B. Web data mining: Exploring hyperlinks, contents, and usage data. Springer-Verlag: Berlin-Heidelberg, 2011. 622 p.
34. Zafarani R., Abbasi M.A. and Liu H. Social media mining. An introduction. Cambridge University Press. 2019. 380 p.
35. Big Data Analysis: New Algorithms for a New Society. N. Japkowicz and J. Stefanowski (eds.), Springer, Switzerland. 2016. 329 p.
36. Data mining for the Internet of things: Literature review and challenges. F. Chen, P. Deng, J. Wan, D. Zhang. *Intern. Journal of Distributed Sensor Networks*. Vol. 2015. 14 p.
37. Esling P. and Agón C. Time-series data mining. *ACM Computing Surveys*. 2012. Vol. 45, Issue 1. P. 12–34.
38. Pearl J. Causality: models, reasoning, and inference. Cambridge: Cambridge Univ. Press. 2000. 526 p.
39. Spirtes P., Glymour C. and Scheines R. Causation, prediction and search. New York: MIT Press, 2001. 543 p.
40. Balabanov O.S. (2017). Knowledge discovery in data and causal models in analytical informatics. Problems in Programming. N. 3. P. 96−112. [in Ukrainian]
41. Peters J., Janzing D. and Schölkopf B. Elements of Causal Inference. Foundations and Learning Algorithms. MIT Press, Cambridge, MA, USA, 2017. 265 p.
42. Shiffrin R.M. Drawing causal inference from Big Data. *Proc. Nat. Acad. Scien. USA*. 2016. Vol. 113, N. 27. P. 7308–7309.
43. Pearl J. and Bareinboim E. External validity: From do-calculus to transportability across populations. *Statistical Science*. 2014. Vol. 29, N 4. P. 579–595.
44. Balabanov O.S. (2011). From covariation to causation. Discovery of structures of dependency in data. System Research and Information Technologies. N. 4. P. 104–118. [In Ukrainian]
45. Balabanov O.S. (2016). Reconstruction of causal networks via analysis of Markov properties. Mathematical Machines and Systems. N. 1. P. 16–26. [In Ukrainian]
46. Giudici P. Financial data science. *Statistics and Probability Letters*. 2018. Vol. 136. P. 160–164.
47. Machine learning. Special issue on applications of machine learning and the knowledge discovery process. R. Kohavi, F. Provost. (Eds.) *Machine Learning*. 1998. Vol. 30, N.2/3. P. 127–274.
48. 22nd SIGKDD Conference on Knowledge Discovery and Data Mining, August 13–17, 2016. San Francisco, California.
49. 24th SIGKDD Conference on Knowledge Discovery and Data Mining, August 19–23, 2018. London, UK.
50. LeCun Y., Bengio Y., Hinton G. Deep learning. *Nature*. 2015. Vol. 521. P. 436–444.
51. Donoho D.L. 50 Years of Data Science. *Journal of Computational and Graphical Statistics*. 2017. Vol. 26, Issue 4. P. 745–766.
52. Bühlmann P. and van de Geer S. Statistics for high-dimensional data: Methods, theory and applications. Springer, 2011. 556 p.
53. Bühlmann P. and van de Geer S. Statistics for big data: A perspective. *Statistics and Probability Letters*. 2018. Vol. 136. P. 37–41.

54. Secchi P. On the role of statistics in the era of big data: A call for a debate. Ibid. P. 10–14.
55. Quarteroni A. The role of statistics in the era of big data: A computational scientist' perspective. Ibid. P. 63–67.
56. Cox D.R., Kartsonaki C., Keogh R.H. Big data: Some statistical issues. Ibid. P. 111–115.
57. James G. M. Statistics within business in the era of big data. Ibid. P. 155–159.
58. Weihs C. and Ickstadt K. Data Science: the impact of statistics. *Intern. Journal of Data Science and Analytics*. 2018. Vol. 6. P. 189–194.
59. Efron B. and Hastie T. Computer age statistical inference. Cambridge University Press, N.Y., 2016. 475 p.
60. Carmichael I. and Marron J.S. Data science vs. statistics: two cultures? *Japanese Journal of Statistics and Data Science*. 2018. Vol. 1, Issue 1. P. 117–138.

Одержано 28.03.2019

Про автора:

Балабанов Олександр Степанович,
доктор фізико-математичних наук,
провідний науковий співробітник.
Кількість наукових публікацій в українських виданнях – 60.
Кількість наукових публікацій в зарубіжних виданнях – 12.
Індекс Хірша – 6.
http://orcid.org/0000-0001-9141-9074.

***Місце роботи автора*:**

Інститут програмних систем
НАН України,
03187, м. Київ-187,
проспект Академіка Глушкова, 40.
Тел.: (044) 5263420.
E-mail: bas@isofts.kiev.ua