Nie wszystkie dane to złoto

Sama ilość przetwarzanych danych nie gwarantuje jeszcze sukcesu w postaci wiarygodnych, wartościowych wyników. Przy rosnących lawinowo zasobach danych sztuką wciąż pozostaje wybór tych odpowiednich i faktycznie przydatnych.

Świat oszalał dzisiaj na punkcie danych. Ich przetwarzanie w systemach informatycznych i automatyczna analiza mają być receptą na wszelkie problemy i wyzwania współczesnego świata. Organizacje i działania sterowane danymi (*data driven*) uznawane są za wzorcowe rozwiązania dla wszelkich niemalże dziedzin życia. Ile w tym racji, a ile marketingowych zaklęć pozostaje kwestią do rozstrzygnięcia. Faktem jednak jest, że na wynikach cyfrowej obróbki danych bazuje dzisiaj coraz więcej form ludzkiej aktywności – od prostych wyborów zakupowych po strategiczne decyzje biznesowe.

Wyobraźnię zarówno opinii publicznej, jak i środowisk technologiczno-biznesowych rozpalają w tej chwili głównie pomysły na nowe, najlepiej od razu przełomowe i innowacyjne możliwości wykorzystania narzędzi analizy danych. Oczekiwania rosną wraz z rozwojem technik sztucznej inteligencji. Oliwy do ognia dolewają przyrastające lawinowo, za sprawą masowego już wykorzystania technik informacyjno-komunikacyjnych, zasoby danych w postaci cyfrowej.

> *W wizjach data driven organisations zasoby danych traktowane są zazwyczaj w sposób całościowy, niezróżnicowany, jak rzecz zastana, pewna, trwała. W tym ujęciu to monolityczny, obiektywnie istniejący surowiec, leżący w zasięgu ręki, gotowy do przetworzenia materiał, który można przerobić na dowolny produkt informacyjny.*

Andrzej Gontarz
ekspert ds. monitoringu rynku w zespole Sektorowej Rady ds. Kompetencji – Informatyka

Czy jednak wszystkie dane są jednakowo dobre do analitycznego zastosowania i równie dobrze nadają się do wykorzystania w algorytmach wspierających grę o rynkową przewagę?

W rzeczywistości efektywne wykorzystanie danych wymaga indywidualnego podejścia do nich i wielu, często żmudnych i długotrwałych zabiegów ich przygotowania. Wszak z tych samych danych można otrzymać zupełnie różne informacje, tak samo jak do uzyskania tej samej informacji można wykorzystać wiele zupełnie różnych danych. Poza tym dane mogą być nieaktualne, niepełne, czy wręcz nawet bezwartościowe dla danego przypadku.

Znana od dawna wszystkim adeptom informatyki zasada GIGO (Garbage In Garbage Out, czyli śmieci na wejściu – śmieci na wyjściu) wciąż obowiązuje. Także w obszarze analityki danych, gdzie jest do dyspozycji coraz więcej technologii

pozwalających na automatyczne czyszczenie i przekształcanie dostępnych danych. – *Nadrzędnym celem jest minimalizacja GIGO: minimalizacja „śmieci", które dostają się do modelu, tak aby model minimalizował liczbę otrzymanych błędnych wyników* – pisze Daniel T. Larose w książce „Odkrywanie wiedzy z danych. Wprowadzenie do eksploracji danych" (PWN 2013).

Ilość nie wystarczy

Jakość danych nabiera szczególnego znaczenia w przypadku technik sztucznej inteligencji, zwłaszcza w obszarze uczenia maszynowego. Algorytmy sztucznej inteligencji potrzebują dostępu do olbrzymiej ilości danych, by uczyć się efektywnie i poprawnie. Specjaliści szacują, że algorytm rozpoznawania obrazów, trenowany na przykład do identyfikacji psa czy kota, potrzebuje „obejrzeć" setki tysięcy, a może nawet i miliony zdjęć, żeby móc z dużym prawdopodobieństwem poprawnie wskazać określone zwierzę.

Kilka lat temu na jednej z konferencji poświęconej sztucznej inteligencji prezentowany był przykład algorytmu trenowanego do rozpoznawania damskich botków. Celem jego stworzenia było m.in. ułatwienie pracy zespołom marketingowym w sklepach internetowych. Po zapoznaniu się z okazałą bazą zdjęć butów, algorytm wskazywał uparcie jako damskie botki buty do gry w piłkę nożną, tzw. korki. Po analizie okazało się, że na zdjęciach udostępnionych algorytmowi do nauki damskie obuwie prezentowane było głównie na tle zielonej trawy. Przykład być może banalny, nie wiadomo też, ile w nim elementów funkcjonującej w każdym środowisku zawodowym legendy. Pokazuje jednak znaczenie doboru odpowiednich danych w pracach nad rozwiązaniami z zakresu sztucznej inteligencji. Dane muszą być odpowiedniej jakości, sama ich ilość nie wystarczy. Algorytm, czyli pewna ściśle określona procedura obliczeniowa, tylko dla właściwych – a nie akurat dostępnych – danych wejściowych daje żądane dane wyjściowe.

– *Z perspektywy uczenia maszynowego rozstrzygająca jest jakość zbiorów, na których uczymy algorytmy. Muszą w reprezentatywny sposób opisywać kontekst rozwiązywanego problemu, np. logi z wyszukiwarki, zawierające informację, które oferty dla jakich fraz były klikane i jakie zajmowały pozycje w wynikach. Często konieczna jest dodatkowa klasyfikacja danych, ręczne tagowanie pod kątem potencjalnego uczenia. To odrębny problem. Skuteczność ostatecznego rozwiązania w krytycznym stopniu zależy od reprezentatywności danych, na których trenowano algorytm* – mówi w wywiadzie zamieszczonym na łamach raportu „AI@Enterprise 2021. Praktyczne zastosowania sztucznej inteligencji w biznesie" Olaf Piotrowski, AI Product Development and Big Data Director w Allegro.

Zasilanie systemów sztucznej inteligencji złymi, niewłaściwymi danymi może prowadzić do tzw. biasu algorytmicznego, inaczej przechyłu czy też skrzywienia algorytmicznego. Jako przykład takiego zjawiska podawany jest często proces rekrutacji do pracy, w którym system analityczny preferuje kandydatów płci męskiej. Podobna sytuacja może mieć miejsce również w przypadku działania algorytmu rozpoznawania twarzy. Jeśli do jego trenowania wykorzystane zostaną zdjęcia osób o jasnej karnacji, to nie będzie on działał skutecznie w przypadku osób czarnoskórych.

Zagrożenie przechyłem algorytmicznym może występować w wielu różnych, także prozaicznych wręcz sytuacjach. Przykładowo, jeśli byśmy chcieli zbadać nastawienie klientów do oferowanych towarów lub usług, to musimy pamiętać, że ci, którzy nabyli już produkt, zazwyczaj wyrażają się o nim pozytywnie (trudno się przyznać do własnego błędu). Należy również pamiętać, że pozytywne opinie w sieci można najzwyczajniej kupić. W internecie pełno jest ogłoszeń firm, które oferują właśnie tego typu usługi.

Nadzór nad jakością

Skąd zatem brać dobre, właściwe dane? Jak zapewnić systemom analitycznym i algorytmom sztucznej inteligencji stały dostęp do odpowiedniej jakości danych? Co lub kto może być gwarantem jakości danych i ich właściwego przygotowania do cyfrowej obróbki?

Wiele firm, szczególnie dużych, radzi sobie wewnętrznie z tym problemem poprzez tworzenie programów *data governance*. Określane są w nich obowiązujące (działy i pracowników na poszczególnych stanowiskach) zasady tworzenia danych i reguły zapewnienia ich jakości, w tym m.in. zakresy odpowiedzialności osób związanych z przetwarzaniem danych. W niektórych przypadkach kwestie te regulowane są w pewnym zakresie przez regulacje prawne (np. Rekomendacja D w bankowości czy obowiązujące wszystkich RODO).

W części przedsiębiorstw zatrudniani są specjalni menedżerowie odpowiedzialni za zarządzanie danymi. Do zadań CDO (*Chief Data Officer*) należy koordynowanie procesu tworzenia danych, ich integracji oraz wykorzystywania w systemach informatycznych. Praca nie jest łatwa, gdyż wymaga zazwyczaj łączenia oczekiwań różnych grup interesów i uwzględniania uwarunkowań funkcjonowania różnych działów w organizacji. Pragmatyka pracy handlowców, księgowych czy marketingowców przekłada się często na różne postrzeganie wymagań odnośnie do charakteru i kształtu potrzebnych im danych. CDO musi umieć łączyć te wszystkie perspektywy, oferując jednocześnie uniwersalne dla całej instytucji rozwiązanie, zapewniające zachowanie odpowiedniej jakości danych na każdym etapie ich przetwarzania.

Firmy nie mogą jednak ograniczać się tylko do wytwarzanych przez siebie danych. Potrzebują również dostępu do ich zewnętrznych zasobów. Jeżeli dane stały się towarem, a zarazem surowcem i paliwem do zasilania algorytmów, to kluczowego znaczenia nabiera kwestia obrotu danymi i jego zasad: jak zapewnić jego kontrolę, aby w obiegu znajdowały się tylko dane dobrej jakości?

Rozwojowy rynek

Według raportu spółki OnAudience, wartość globalnego rynku danych rosła w latach 2017–22 w tempie 29 proc. rocznie i wynosi obecnie 52 mld USD. Dominującą pozycję mają na nim Stany Zjednoczone. Dla 52 proc. uczestników badania „State of the CIO, 2022" analityka danych i uczenie maszynowe będą w tym roku w ich firmach obszarami największych inwestycji w informatykę.

Chiński eksperyment

Warto przyjrzeć się pomysłowi testowanemu w Chinach. W listopadzie ub.r., po 5 latach przygotowań, rozpoczęła działalność Szanghajska Giełda Danych (*Shanghai Data Exchange*). Chiny chcą przetestować możliwości handlu danymi na wzór obrotu innymi produktami i towarami. Doświadczenia z tego programu pilotażowego mają posłużyć do uruchomienia kolejnych giełd i stworzenia krajowego, a także globalnego rynku handlu danymi. Otwarcie giełdy połączone było z obradami „Global Data Ecosystem Conference 2021".

W początkowym okresie w ofercie szanghajskiej giełdy znalazło się 100 produktów związanych z danymi. Są one dostarczane m.in. przez takie firmy, jak: China Eastern Airlines, China Mobile, China Unicom, China Telecom, Cosco Shipping, Sinofaith, State Grid Shanghai Electric Power, ICBC. Pogrupowano je w osiem kategorii, m.in. finanse, komunikacja, transport. W miarę rozwoju giełdy dostępnych produktów będzie przybywać.

Obrót danymi odbywa się pod nadzorem komitetu ekspertów, w skład którego wchodzi 31 specjalistów z zakresu: zgodności z prawem (*compliance*), transakcji finansowych czy bezpieczeństwa danych. Korzystającym z giełdy zapewnia się przejrzystość transakcji – wiadomo, kto komu jakie dane sprzedaje.

– Giełda ma na celu rozwiązanie pięciu najważniejszych kwestii związanych z handlem danymi: weryfikacji praw podmiotu do sprzedaży danych, ich wyceny, wzajemnego zaufania obu stron transakcji, wejścia na rynek i nadzoru. Giełda oferuje usługi przedsprzedażowe (np. ocena jakości pakietów danych), jak i posprzedażowe (np. weryfikacja transakcji, arbitraż) – tłumaczy na łamach magazynu CRN (nr 1/2022) Albert Borowiecki. Zwraca przy tym uwagę, że rynek danych regulują w Chinach trzy ustawy: Cyber Security Law, Data Security Law oraz Personal Information Protection Law.

Perspektywa społeczna

I jeszcze jedna kwestia, która w dyskusjach o warunkach efektywnego wykorzystania danych nie jest jeszcze pierwszoplanowa, ale dla oceny przydatności danych może mieć kluczowe znaczenie. Chodzi o kulturowy kontekst wytwarzania danych i ich zastosowania. Dane wytwarzane są zazwyczaj w trakcie aktywności człowieka i poddawane są takim samym wpływom środowiska społecznego, jak wszystkie inne ludzkie wytwory (można dyskutować, czy dane przekazywane przez urządzenie IoT nie mają również społecznego wymiaru, chociażby przez sam fakt decydowania przez człowieka o takim, a nie innym ich wykorzystaniu).

Kwestie te nabierają szczególnego znaczenia w przypadku wystąpienia wspomnianego zjawiska przechyłu algorytmicznego. Już toczą się dyskusje, na ile można w takiej sytuacji ingerować w dane, jeżeli faktycznie odzwierciedlają one istniejący społecznie stan rzeczy (np. pracodawcy rzeczywiście preferują kandydatów mężczyzn), chociaż nie odpowiada on aktualnym potrzebom czy oczekiwaniom twórców lub użytkowników algorytmów. W toczonych dyskusjach pojawiają się już nawet przestrogi przed zagrożeniem cenzurowaniem danych. Proponowanym wyjściem z sytuacji mogłoby być, zdaniem niektórych, pozwolenie algorytmowi na dłuższe uczenie się, z użyciem nowych, pojawiających w praktyce społecznej danych, odzwierciedlających nowe, akceptowane społecznie postawy i zachowania.

Dostęp do odpowiednich danych i dzielenie się nimi mogą umożliwiać firmom nowoczesne rozwiązania technologiczne, na przykład dedykowane platformy sieciowe do wymiany danych. Twórcy raportu „Trendy technologiczne 2022" z firmy Deloitte zwracają uwagę na dokonujący się postęp w zakresie bezpieczeństwa danych, co może z kolei stymulować rozwój bazującej na wykorzystaniu danych współpracy nawet w sytuacji konkurowania firm między sobą na rynku. – *Po raz pierwszy pojawia się możliwość przesłania zaszyfrowanych wrażliwych danych do celów analitycznych – takich, które tradycyjnie były absolutnie niedostępne dla zewnętrznych podmiotów. Ten trend stwarza nowe możliwości w zakresie monetyzacji danych, a także umożliwia współpracę pomiędzy konkurentami na niespotykaną wcześniej skalę* – czytamy na stronie Deloitte'a.

Ciekawym rozwiązaniem, wskazywanym w raporcie, mogą być też tzw. obliczenia wielostronne, które umożliwiają rozłożenie analizy danych na wiele podmiotów, przy czym żaden z nich osobno nie może zobaczyć pełnego zestawu informacji wejściowych. To stwarza m.in. możliwości wykorzystania wartościowych zasobów danych w sytuacji, gdy ich posiadacz – upatrując w nich przewagi konkurencyjnej – nie chce ich ujawniać.