Definiowanie roli danych

Nie ma nic nowego w danych. Każda ciekawa aplikacja napisana dla komputera ma powiązane z nią dane. Dane mają różne formy - niektóre są uporządkowane, inne nie. Zmieniła się ilość danych. Niektórzy uważają za przerażające, że mamy teraz dostęp do tak wielu danych, które opisują prawie każdy aspekt życia większości ludzi, czasem nawet do poziomu, którego nawet ta osoba nie zdaje sobie sprawy. Ponadto zastosowanie zaawansowanego sprzętu i ulepszenia algorytmów sprawiają, że dane są dziś uniwersalnym zasobem sztucznej inteligencji. Aby pracować z danymi, musisz je najpierw uzyskać. Obecnie aplikacje zbierają dane ręcznie, jak miało to miejsce w przeszłości, a także automatycznie, przy użyciu nowych metod. Nie jest to jednak tylko jedna lub dwie techniki gromadzenia danych; metody zbierania odbywają się w sposób ciągły od manualnego do w pełni automatycznego. Surowe dane zwykle nie działają dobrze do celów analizy. Ten rozdział pomaga także zrozumieć potrzebę manipulowania danymi i ich kształtowania, tak aby spełniały określone wymagania. Odkrywasz także potrzebę zdefiniowania prawdziwej wartości danych, aby przede wszystkim upewnić się, że wyniki analizy odpowiadają celom ustalonym dla aplikacji. Co ciekawe, masz również do czynienia z limitami gromadzenia danych. Obecnie nie istnieje technologia umożliwiająca wyłapywanie myśli z czyjegoś umysłu za pomocą środków telepatycznych. Oczywiście istnieją również inne ograniczenia - z których większość prawdopodobnie już wiesz, ale mogłeś nie wziąć pod uwagę.

Znalezienie danych wszechobecnych w tym wieku

Rewolucja w zakresie dużych zbiorów danych jest czymś więcej niż modnym hasłem używanym przez dostawców do proponowania nowych sposobów przechowywania i analizowania danych, jest codzienną rzeczywistością i siłą napędową naszych czasów. Być może słyszałeś o dużych danych wspomnianych w wielu specjalistycznych publikacjach naukowych i biznesowych, a nawet zastanawiałeś się, co tak naprawdę oznacza ten termin. Z technicznego punktu widzenia duże dane odnoszą się do dużych i złożonych ilości danych komputerowych, tak dużych i skomplikowanych, że aplikacje nie mogą sobie z nimi poradzić, wykorzystując dodatkową pamięć lub zwiększając moc komputera. Big Data oznacza rewolucję w przechowywaniu danych i manipulowaniu nimi. Wpływa na to, co można osiągnąć dzięki danym w sposób bardziej jakościowy (oprócz wykonywania większej liczby zadań można lepiej wykonywać zadania). Komputery przechowują duże dane w różnych formatach z ludzkiej perspektywy, ale komputer traktuje dane jako strumień zer i jedynek (podstawowy język komputerów). Dane można wyświetlać jako jeden z dwóch typów, w zależności od sposobu ich produkcji i zużycia. Niektóre dane mają przejrzystą strukturę (dokładnie wiesz, co zawiera i gdzie znaleźć każdy kawałek danych), podczas gdy inne dane są nieustrukturyzowane (masz pojęcie o tym, co zawiera, ale nie wiesz dokładnie, jak się je ułoży) . Typowymi przykładami danych strukturalnych są tabele bazy danych, w których informacje są uporządkowane w kolumnach, a każda kolumna zawiera określony typ informacji. Dane są często uporządkowane według projektu. Zbierasz je selektywnie i zapisujesz we właściwym miejscu. Na przykład możesz chcieć umieścić liczbę osób kupujących określony produkt w określonej kolumnie, w określonej tabeli, w określonej bazie danych. Podobnie jak w przypadku biblioteki, jeśli wiesz, jakich danych potrzebujesz, możesz je natychmiast znaleźć. Nieustrukturyzowane dane składają się z obrazów, filmów i nagrań dźwiękowych. Możesz użyć nieustrukturyzowanego formularza dla tekstu, aby oznaczyć go takimi cechami, jak rozmiar, data lub typ zawartości. Zazwyczaj nie wiesz dokładnie, gdzie dane pojawiają się w nieustrukturyzowanym zbiorze danych, ponieważ dane pojawiają się jako sekwencje zer i jedynek, które aplikacja musi interpretować lub wizualizować.

Przekształcanie nieustrukturyzowanych danych w ustrukturyzowaną formę może kosztować dużo czasu i wysiłku oraz może wymagać pracy wielu osób. Większość danych z rewolucji dużych zbiorów danych jest nieuporządkowana i przechowywana w obecnej postaci, chyba że ktoś to ustrukturyzuje

Ten obszerny i wyrafinowany magazyn danych nie pojawił się nagle z dnia na dzień. Opracowanie technologii przechowywania tej ilości danych wymagało czasu. Ponadto upowszechnienie technologii generowania i dostarczania danych wymagało czasu, a mianowicie komputerów, czujników, inteligentnych telefonów komórkowych, Internetu i usług w sieci WWW. Poniższe sekcje pomagają zrozumieć, co sprawia, że dane są dziś uniwersalnym zasobem.

Zrozumienie implikacji Moore'a

W 1965 r. Gordon Moore, współzałożyciel Intela i Fairchild Semiconductor, napisał w artykule zatytułowanym „Wbijanie większej liczby komponentów w układy scalone", że liczba elementów w układach scalonych podwoi się każdego roku w ciągu następnej dekady. W tym czasie tranzystory zdominowały elektronikę. Możliwość umieszczenia większej liczby tranzystorów w układzie scalonym (IC) oznaczała możliwość uczynienia urządzeń elektronicznych bardziej wydajnymi i użytecznymi. Proces ten nazywa się integracją i implikuje silny proces miniaturyzacji elektroniki (znacznie zmniejszając ten sam obwód). Dzisiejsze komputery nie są wcale mniejsze od komputerów sprzed dekady, ale są zdecydowanie mocniejsze. To samo dotyczy telefonów komórkowych. Mimo że są tego samego rozmiaru co ich poprzednicy, są w stanie wykonywać więcej zadań. To, co Moore stwierdził w tym artykule, było prawdą przez wiele lat. Przemysł półprzewodników nazywa to, jak przewidywano, podwojeniem prawa Moore'a przez pierwsze dziesięć lat. W 1975 r. Moore poprawił swoje oświadczenie, przewidując podwojenie co dwa lata. Ten wskaźnik podwojenia jest nadal aktualny, choć obecnie powszechnie uważa się, że nie utrzyma się on dłużej niż do końca obecnej dekady (do około 2020 r.). Począwszy od 2012 r. zaczęło występować niedopasowanie między oczekiwanym wzrostem prędkości a tym, co firmy półprzewodnikowe mogą osiągnąć w zakresie miniaturyzacji. Istnieją fizyczne przeszkody w integracji większej liczby obwodów w układzie scalonym z wykorzystaniem obecnych składników krzemionki, ponieważ można uczynić rzeczy tak małymi. Jednak innowacje trwają. W przyszłości prawo Moore'a może nie mieć zastosowania, ponieważ przemysł przejdzie na nową technologię (na przykład wytwarzanie komponentów za pomocą laserów optycznych zamiast tranzystorów). Ważne jest to, że od 1965 roku podwojenie komponentów co dwa lata zapoczątkowało wielki postęp w elektronice cyfrowej, co miało dalekosiężne konsekwencje w pozyskiwaniu, przechowywaniu, manipulowaniu i zarządzaniu danymi. Prawo Moore'a ma bezpośredni wpływ na dane. Zaczyna się od inteligentniejszych urządzeń. Im inteligentniejsze urządzenia, tym większa dyfuzja (o czym świadczy dziś elektronika). Im większa dyfuzja, tym niższa staje się cena, tworząc nieskończoną pętlę, która napędza użycie potężnych komputerów i małych czujników na całym świecie. Przy dużej ilości dostępnej pamięci komputera i większych dyskach do przechowywania danych konsekwencją jest zwiększenie dostępności danych, takich jak strony internetowe, zapisy transakcji, pomiary, obrazy cyfrowe i inne rodzaje danych.

Używanie danych wszędzie

Naukowcy potrzebują silniejszych komputerów niż przeciętny człowiek ze względu na swoje eksperymenty naukowe. Zaczęli radzić sobie z imponującą ilością danych na wiele lat, zanim ktokolwiek wymyślił pojęcie dużych zbiorów danych. W tym momencie Internet nie wygenerował ogromnych ilości danych, które robi dzisiaj. Pamiętaj, że duże zbiory danych nie są modą tworzoną przez dostawców oprogramowania i sprzętu, ale mają podstawy w wielu dziedzinach naukowych, takich jak astronomia (misje kosmiczne), satelita (nadzór i monitorowanie), meteorologia, fizyka (akceleratory cząstek) i genomika (sekwencje DNA). Chociaż aplikacje AI mogą specjalizować się w dziedzinie naukowej, takiej jak Watson firmy IBM, która ma imponującą zdolność diagnozowania medycznego, ponieważ może uczyć się informacji z milionów artykułów naukowych na temat chorób i medycyny, rzeczywisty sterownik aplikacji AI ma często bardziej przyziemne aspekty. Rzeczywiste aplikacje AI są w większości cenione za to, że potrafią rozpoznawać obiekty, poruszać się po ścieżkach lub rozumieć, co ludzie

mówią i do nich. Wkład danych w renesans AI, który ukształtował go w taki sposób, nie pochodził z klasycznych źródeł danych naukowych. Internet generuje i dystrybuuje nowe dane w dużych ilościach. Nasza obecna dzienna produkcja danych szacowana jest na około 2,5 kwintyliona (liczba z 18 zerami) bajtów, przy czym lwia część będzie przeznaczona na nieustrukturyzowane dane, takie jak filmy i pliki audio. Wszystkie te dane są związane z powszechnymi ludzkimi działaniami, uczuciami, doświadczeniami i relacjami. Wędrując po tych danych, sztuczna inteligencja może łatwo nauczyć się, jak działa rozumowanie i działanie bardziej ludzkie. Oto kilka przykładów bardziej interesujących danych, które można znaleźć:

* Duże repozytoria twarzy i wyrazów ze zdjęć i filmów opublikowane w serwisach społecznościowych, takich jak Facebook, YouTube i Google, zawierają informacje o płci, wieku, uczuciach i ewentualnie preferencjach seksualnych, orientacjach politycznych lub IQ

* Prywatne informacje medyczne i dane biometryczne z inteligentnych zegarków, które mierzą dane ciała, takie jak temperatura i tętno podczas choroby i zdrowia.

* Zestawy danych o relacjach między ludźmi i ich zainteresowaniach ze źródeł takich jak media społecznościowe i wyszukiwarki. Na przykład badanie z Centrum Psychometrii Uniwersytetu Cambridge twierdzi, że interakcje na Facebooku zawierają wiele danych o relacjach intymnych ~

* Informacje o tym, jak mówimy, są rejestrowane przez telefony komórkowe. Na przykład OK Google, funkcja dostępna w telefonach z Androidem, rutynowo zapisuje pytania, a czasem nawet więcej

Każdego dnia użytkownicy podłączają do Internetu jeszcze więcej urządzeń, które zaczynają przechowywać nowe dane osobowe. Obecnie w domach są asystenci, tacy jak Amazon Echo i inne zintegrowane urządzenia inteligentnego domu, które oferują sposoby regulowania i ułatwiania warunków domowych. To tylko wierzchołek góry lodowej, ponieważ wiele innych powszechnych narzędzi codziennego życia łączy się (od lodówki do szczoteczki do zębów) i może przetwarzać, rejestrować i przesyłać dane. Internet przedmiotów (IoT) staje się rzeczywistością. Eksperci szacują, że do 2020 r. Będzie istniało sześć razy więcej powiązanych rzeczy niż osób, ale zespoły badawcze i ośrodki analityczne już odwiedzają te dane.

Wprowadzanie do działania algorytmów

Rasa ludzka znajduje się obecnie na niesamowitym skrzyżowaniu niespotykanych dotąd ilości danych, generowanych przez coraz mniejszy i potężniejszy sprzęt. Dane są również coraz częściej przetwarzane i analizowane przez te same komputery, na których proces pomógł rozprzestrzeniać się i rozwijać. To stwierdzenie może wydawać się oczywiste, ale dane stały się tak wszechobecne, że ich wartość nie zależy już tylko od zawartych w nich informacji (takich jak przypadek danych przechowywanych w bazie danych firmy, która umożliwia jej codzienne operacje), ale raczej jako wykorzystanie oznacza tworzenie nowych wartości; takie dane są określane jako „nowy olej". Te nowe wartości istnieją głównie w sposobie, w jaki aplikacje robią manicure, przechowują i pobierają dane oraz w jaki sposób faktycznie je wykorzystujesz za pomocą inteligentnych algorytmów. Algorytmy i sztuczna inteligencja zmieniły grę danych. Jak już wspomniano , algorytmy AI próbowały po drodze różnych podejść, przechodząc od prostych algorytmów do symbolicznego wnioskowania opartego na logice, a następnie do systemów eksperckich. W ostatnich latach stały się sieciami neuronowymi i, w swojej najbardziej dojrzałej formie, głębokim uczeniem się. W trakcie tego przejścia metodologicznego dane przekształciły się z informacji przetwarzanych przez z góry określone algorytmy w przekształcenia algorytmu w coś przydatnego do tego zadania. Dane przekształciły się z samego surowca, który napędzał rozwiązanie, w rzemieślnika samego rozwiązania. Dlatego zdjęcie niektórych twoich kociąt staje się coraz bardziej przydatne nie tylko ze względu na ich wartość afektywną - przedstawiającą twoje słodkie małe kotki - ale także

dlatego, że może stać się częścią procesu uczenia się sztucznej inteligencji odkrywającej bardziej ogólne pojęcia, takie jak jakie cechy oznacz kota lub zrozumienie, co definiuje słodkie. Na większą skalę firma taka jak Google pobiera swoje algorytmy z swobodnie dostępnych danych, takich jak treść stron internetowych lub tekst znajdujący się w publicznie dostępnych tekstach i książkach. Oprogramowanie pająka Google indeksuje sieć, przeskakuje z witryny na witrynę, pobierając strony internetowe z zawartością tekstu i obrazów. Nawet jeśli Google oddaje część danych użytkownikom jako wyniki wyszukiwania, wyodrębnia inne dane z danych za pomocą algorytmów AI, które uczą się od niej, jak osiągnąć inne cele. Algorytmy przetwarzające słowa mogą pomóc systemom AI w Google zrozumieć i przewidzieć Twoje potrzeby, nawet jeśli nie wyrażasz ich w zestawie słów kluczowych, ale w prostym, niejasnym języku naturalnym, którym mówimy na co dzień (i tak, język codzienny jest często niejasny) . Jeśli obecnie próbujesz zadać wyszukiwarce Google pytania, a nie tylko łańcuchy słów kluczowych, zauważysz, że odpowiedź jest poprawna. Od 2012 roku, wraz z wprowadzeniem aktualizacji Hummingbird, Google lepiej rozumie synonimy i pojęcia, co wykracza poza początkowe dane, które uzyskał, i to jest wynik procesu AI. W Google istnieje jeszcze bardziej zaawansowany algorytm o nazwie RankBrain, który uczy się bezpośrednio z milionów zapytań każdego dnia i może odpowiadać na dwuznaczne lub niejasne zapytania wyszukiwania, nawet wyrażone w wyrażeniach slangowych lub potocznych lub po prostu z błędami. RankBrain nie obsługuje wszystkich zapytań, ale uczy się na podstawie danych, jak lepiej odpowiadać na zapytania. Obsługuje już 15 procent zapytań wyszukiwarki, a w przyszłości odsetek ten może wynieść 100 procent.

Pomyślne wykorzystanie danych

Dostęp do dużej ilości danych nie wystarczy, aby stworzyć udaną sztuczną inteligencję. Obecnie algorytm AI nie może wyodrębniać informacji bezpośrednio z surowych danych. Większość algorytmów opiera się na zewnętrznym zbiorze i manipulacji przed analizą. Gdy algorytm zbiera przydatne informacje, może nie reprezentować właściwych informacji. Poniższe sekcje pomagają zrozumieć, jak gromadzić, manipulować i automatyzować gromadzenie danych z perspektywy ogólnej.

Biorąc pod uwagę źródła danych

Wykorzystywane dane pochodzą z wielu źródeł. Najczęstsze źródło danych pochodzi z informacji wprowadzonych przez ludzi w pewnym momencie. Nawet gdy system automatycznie zbiera dane z witryn zakupów, ludzie początkowo wprowadzają te informacje. Człowiek klika różne przedmioty, dodaje je do koszyka, określa cechy (takie jak rozmiar) i ilość, a następnie sprawdza. Później, po sprzedaży, człowiek ocenia ocenę zakupów, produkt i sposób dostawy i komentuje. Krótko mówiąc, każde doświadczenie zakupowe staje się również ćwiczeniem polegającym na gromadzeniu danych. Wiele źródeł danych opiera się obecnie na danych zebranych ze źródeł ludzkich. Ludzie zapewniają również ręczne wprowadzanie. Dzwonisz lub udajesz się do jakiegoś biura, aby umówić się z profesjonalistą. Recepcjonista zbiera następnie informacje potrzebne do spotkania. Te ręcznie zebrane dane ostatecznie kończą się gdzieś w zbiorze danych do celów analizy. Dane są również gromadzone z czujników i czujniki te mogą przybierać prawie dowolną formę. Na przykład wiele organizacji opiera wykrywanie danych fizycznych, takich jak liczba osób oglądających obiekt w oknie, na wykryciu telefonu komórkowego. Oprogramowanie do rozpoznawania twarzy może potencjalnie wykryć powtarzających się klientów. Jednak czujniki mogą tworzyć zestawy danych z prawie wszystkiego. Usługa pogodowa opiera się na zestawach danych utworzonych przez czujniki monitorujące warunki środowiskowe, takie jak deszcz, temperatura, wilgotność, zachmurzenie i tak dalej. Zrobotyzowane systemy monitorowania pomagają korygować niewielkie wady w robotycznym działaniu poprzez ciągłą analizę danych gromadzonych przez czujniki monitorujące. Czujnik w połączeniu z niewielką aplikacją AI może powiedzieć, kiedy obiad zostanie ugotowany do perfekcji dziś wieczorem. Czujnik zbiera dane, ale aplikacja AI używa reguł, które pomagają określić, kiedy jedzenie jest odpowiednio ugotowane.

Uzyskiwanie wiarygodnych danych

Słowo „wiarygodny" wydaje się tak łatwe do zdefiniowania, ale jednocześnie trudne do wdrożenia. Coś jest wiarygodne, gdy uzyskane wyniki są oczekiwane i spójne. Wiarygodne źródło danych wytwarza przyziemne dane, które nie zawierają niespodzianek; nikt nie jest zaskoczony rezultatem. W zależności od Twojej perspektywy może być dobrą rzeczą, że większość ludzi nie ziewa, a potem nie śpi podczas przeglądania danych. Niespodzianki sprawiają, że dane warte są analizy i przeglądu. W związku z tym dane mają aspekt dualności. Chcemy wiarygodnych, przyziemnych, w pełni przewidywanych danych, które po prostu potwierdzają to, co już wiemy, ale nieoczekiwane jest to, co sprawia, że zbieranie danych jest przydatne przede wszystkim. Nadal nie chcesz danych, które są tak niezwykłe, że ich przeglądanie staje się niemal przerażające. Podczas uzyskiwania danych należy zachować równowagę. Dane muszą mieścić się w określonych granicach (zgodnie z opisem w części „Zarządzanie danymi", w dalszej części tego rozdziału). Musi również spełniać określone kryteria co do wartości prawdy (jak opisano w sekcji „Uwzględnianie pięciu niepoprawności w danych", w dalszej części tego rozdziału). Dane muszą także przychodzić w oczekiwanych odstępach czasu, a wszystkie pola rekordu danych przychodzących muszą być kompletne.

Do pewnego stopnia bezpieczeństwo danych wpływa również na niezawodność danych. Spójność danych występuje w kilku formach. Gdy dane dotrą, możesz upewnić się, że mieszczą się one w oczekiwanych zakresach i pojawiają się w określonej formie. Jednak po zapisaniu danych niezawodność może się zmniejszyć, chyba że upewnisz się, że dane pozostaną w oczekiwanej formie. Podmiot majstrujący przy danych wpływa na niezawodność, czyniąc dane podejrzanymi i potencjalnie niezdatnymi do późniejszej analizy. Zapewnienie niezawodności danych oznacza, że po dostarczeniu danych nikt nie manipuluje nim, aby dopasować go do oczekiwanej domeny (w rezultacie czyniąc go prozaicznym).

Zwiększanie niezawodności wkładu człowieka

Ludzie popełniają błędy - to część bycia człowiekiem. W rzeczywistości oczekiwanie, że ludzie nie popełnią błędów, jest nieuzasadnione. Jednak wiele projektów zakłada, że ludzie w jakiś sposób nie popełniają błędów. Projekt oczekuje, że wszyscy będą przestrzegać zasad. Niestety zdecydowana większość użytkowników ma gwarancję, że nawet nie przeczyta reguł, ponieważ większość ludzi jest też leniwa lub zbyt naciska na czas, jeśli chodzi o robienie rzeczy, które tak naprawdę nie pomagają im bezpośrednio. Rozważ wprowadzenie stanu do formularza. Jeśli podasz tylko pole tekstowe, niektórzy użytkownicy mogą wprowadzić całą nazwę stanu, na przykład Kansas. Oczywiście, niektórzy użytkownicy popełniają literówkę lub błąd wielkich liter i wymyślą Kansus lub kANSAS. Ustawiając te błędy, ludzie i organizacje mają różne podejścia do wykonywania zadań. Ktoś z branży wydawniczej może skorzystać z przewodnika po stylu Associated Press (AP) i wprowadzić Kan. Ktoś, kto jest starszy i przyzwyczajony do wytycznych Rządowego Biura Druku (GPO), może wprowadzić Kans. zamiast. Używane są również inne skróty. U.S. Post Office (USPS) używa KS, ale US Coast Guard używa KA. Tymczasem formularz Międzynarodowej Organizacji Normalizacyjnej (ISO) jest zgodny z US-KS. Pamiętaj, że jest to tylko wpis stanu, który jest dość prosty - a przynajmniej tak myślałeś przed przeczytaniem. Oczywiście, ponieważ stan nie zmieni nazwy w najbliższym czasie, możesz po prostu podać w formularzu listę rozwijaną do wyboru stanu w wymaganym formacie, eliminując w ten sposób różnice w użyciu skrótów, literówek i błędów pisowni w jeden upadek.

Rozwijane pola listy działają dobrze dla niesamowitej gamy danych wejściowych, a ich użycie zapewnia, że dane wejściowe człowieka w tych polach stają się niezwykle niezawodne, ponieważ człowiek nie ma wyboru, jak tylko użyć jednej z domyślnych pozycji. Oczywiście człowiek zawsze może wybrać niewłaściwy wpis, w którym wchodzą podwójne kontrole. Niektóre nowsze aplikacje porównują kod

pocztowy z miastem i wpisem stanu, aby sprawdzić, czy są zgodne. Gdy się nie zgadzają, użytkownik jest ponownie proszony o podanie poprawnych danych wejściowych. To podwójne sprawdzenie graniczy z irytacją (szczegółowe informacje można znaleźć na pasku bocznym „Bardziej irytujące niż użyteczne pomoce wejściowe"), ale użytkownik raczej nie zobaczy go zbyt często, więc nie powinno być zbyt denerwujące.

Nawet przy kontroli krzyżowej i zapisach statycznych ludzie nadal mają dużo miejsca na popełnianie błędów. Na przykład wprowadzanie liczb może być problematyczne. Gdy użytkownik musi wprowadzić 2,00, możesz zobaczyć 2, 2.0, 2. lub dowolną z wielu innych pozycji. Na szczęście parsowanie wpisu i jego ponowne sformatowanie rozwiąże problem. Możesz wykonać to zadanie automatycznie, bez pomocy użytkownika. Niestety ponowne formatowanie nie poprawi błędnych danych liczbowych. Możesz częściowo ograniczyć takie błędy, włączając sprawdzanie zasięgu. Klient nie może kupić -5 kostek mydła. Uzasadnionym sposobem pokazania klientowi zwrotu kostek mydła jest przetworzenie zwrotu, a nie sprzedaży. Jednak użytkownik mógł po prostu popełnić błąd i możesz podać komunikat określający odpowiedni zakres wejściowy dla wartości.

Korzystanie z automatycznego gromadzenia danych

Niektóre osoby uważają, że automatyczne gromadzenie danych rozwiązuje wszystkie problemy związane z wprowadzaniem danych przez człowieka związane z zestawami danych. W rzeczywistości automatyczne zbieranie danych zapewnia szereg korzyści:

* Lepsza spójność

* Poprawiona niezawodność

* Niższe prawdopodobieństwo braku danych

* Zwiększona dokładność

* Zmniejszona wariancja dla rzeczy takich jak wejścia czasowe

Niestety stwierdzenie, że automatyczne gromadzenie danych rozwiązuje każdy problem, jest po prostu nieprawidłowe. Zautomatyzowane zbieranie danych nadal opiera się na czujnikach, aplikacjach i sprzęcie komputerowym zaprojektowanym przez ludzi, które zapewniają dostęp tylko do danych, na które ludzie zezwalają. Ze względu na ograniczenia, jakie ludzie nakładają na cechy automatycznego gromadzenia danych, wynik często dostarcza mniej pomocnych informacji, niż oczekiwali projektanci. W związku z tym automatyczne gromadzenie danych podlega ciągłym zmianom, gdy projektanci próbują rozwiązać problemy z danymi wejściowymi. Zautomatyzowane zbieranie danych ma również błędy programowe i sprzętowe występujące w dowolnym systemie komputerowym, ale ma większy potencjał do rozwiązywania problemów miękkich (które powstają, gdy system najwyraźniej działa, ale nie zapewnia pożądanych rezultatów) niż inne rodzaje komputerów ustawienia. Kiedy system działa, niezawodność danych wejściowych znacznie przekracza ludzkie możliwości. Jednak w przypadku wystąpienia problemów miękkich system często nie rozpoznaje problemu, tak jak istniałby człowiek, dlatego zestaw danych może zawierać bardziej przeciętne lub nawet złe dane.

Manicuring the Data

Niektóre osoby używają terminu manipulacja, mówiąc o danych, sprawiając wrażenie, że dane są w jakiś sposób zmieniane w sposób pozbawiony skrupułów lub przebiegły. Być może lepszym terminem byłoby manicure, co sprawia, że dane są dobrze ukształtowane i piękne. Niezależnie od tego, jakiego terminu używasz, surowe dane rzadko spełniają wymagania dotyczące przetwarzania i analizy. Aby

uzyskać coś z danych, musisz to zrobić, aby spełnić określone potrzeby. W poniższych sekcjach omówiono potrzeby związane z obsługą danych.

Radzenie sobie z brakującymi danymi

Aby poprawnie odpowiedzieć na dane pytanie, musisz mieć wszystkie fakty. Możesz odgadnąć odpowiedź na pytanie bez wszystkich faktów, ale odpowiedź jest równie błędna, jak poprawna. Często mówi się, że ktoś, kto podejmuje decyzję, zasadniczo odpowiadając na pytanie, bez wszystkich faktów, doszedł do wniosku. Analizując dane, prawdopodobnie wyciągnąłeś więcej wniosków niż myślisz z powodu brakujących danych. Rekord danych, jeden wpis w zestawie danych (który jest wszystkimi danymi), składa się z pól zawierających fakty użyte do udzielenia odpowiedzi na pytanie. Każde pole zawiera jeden rodzaj danych, które dotyczą jednego faktu. Jeśli to pole jest puste, nie masz danych potrzebnych do odpowiedzi na pytanie przy użyciu tego konkretnego rekordu danych.

W ramach postępowania z brakującymi danymi musisz wiedzieć, że brakuje danych. Stwierdzenie, że w zestawie danych brakuje informacji, może być naprawdę trudne, ponieważ wymaga spojrzenia na niskim poziomie danych - coś, na co większość ludzi nie jest przygotowana i zajmuje dużo czasu, nawet jeśli masz wymagane umiejętności. Często pierwszą wskazówką, że brakuje danych, jest niedorzeczna odpowiedź na twoje pytanie z algorytmu i powiązanego zestawu danych. Gdy algorytm jest właściwy do użycia, zestaw danych musi być uszkodzony. Problem może wystąpić, gdy proces gromadzenia danych nie obejmuje wszystkich danych potrzebnych do udzielenia odpowiedzi na określone pytanie. Czasami lepiej jest zrezygnować z faktu, a nie skorzystać z poważnie uszkodzonego faktu. Jeśli okaże się, że w określonym polu w zbiorze danych brakuje 90 procent lub więcej jego danych, pole staje się bezużyteczne i należy je usunąć z zestawu danych (lub znajdź jakiś sposób na uzyskanie wszystkich tych danych). W mniej uszkodzonych polach może brakować danych na jeden z dwóch sposobów. Przypadkowo brakujące dane są często wynikiem błędu człowieka lub czujnika. Występuje, gdy w rekordach danych w zestawie danych brakuje wpisów. Czasami zwykła usterka spowoduje uszkodzenie. Sekwencyjnie brakujące dane występują podczas pewnego rodzaju awarii uogólnionej. Cały segment rekordów danych w zbiorze danych nie ma wymaganych informacji, co oznacza, że wynikowa analiza może być dość wypaczona. Naprawianie losowo brakujących danych jest najłatwiejsze. Jako zamiennika można użyć prostej wartości średniej lub średniej. Nie, zestaw danych nie jest całkowicie dokładny, ale prawdopodobnie będzie działać wystarczająco dobrze, aby uzyskać rozsądną odpowiedź. W niektórych przypadkach badacze danych używali specjalnego algorytmu do obliczania brakującej wartości, co może zwiększyć dokładność zestawu danych kosztem czasu obliczeniowego. Sekwencyjnie brakujące dane są znacznie trudniejsze, jeśli nie niemożliwe, do naprawienia, ponieważ brakuje Ci danych otaczających, na których można by się domyślić. Jeśli znajdziesz przyczynę brakujących danych, czasami możesz je zrekonstruować. Jednak gdy odbudowa stanie się niemożliwa, możesz zignorować pole. Niestety niektóre odpowiedzi będą wymagały tego pola, co oznacza, że może być konieczne zignorowanie tej konkretnej sekwencji rekordów danych - potencjalnie powodując nieprawidłowe dane wyjściowe. Uwzględnianie niedopasowania danych Dane mogą istnieć dla każdego rekordu danych w zestawie danych, ale mogą nie być wyrównywane z innymi danymi w innych posiadanych zestawach danych. Na przykład dane liczbowe w polu w jednym zbiorze danych mogą być typu zmiennoprzecinkowego (z kropką dziesiętną), ale typem całkowitym w innym zbiorze danych. Przed połączeniem dwóch zestawów danych pola muszą zawierać ten sam typ danych. Mogą wystąpić wszelkiego rodzaju inne rodzaje niewspółosiowości. Na przykład pola daty są znane z tego, że są formatowane na różne sposoby. Aby porównać daty, formaty danych muszą być takie same. Jednak daty są również podstępne w ich skłonności do wyglądania tak samo, ale nie są takie same. Na przykład daty w jednym zestawie danych mogą wykorzystywać jako podstawę czas GMT (Greenwich Mean Time), podczas gdy daty w innym zestawie danych mogą wykorzystywać inną strefę czasową.

Aby porównać czasy, musisz je wyrównać w tej samej strefie czasowej. Może stać się jeszcze dziwniejsze, gdy daty w jednym zestawie danych pochodzą z lokalizacji, która korzysta z czasu letniego (DST), ale daty z innej lokalizacji nie.

Nawet jeśli typy danych i format są takie same, mogą wystąpić inne niedopasowania danych. Na przykład pola w jednym zestawie danych mogą nie pasować do pól w drugim zestawie danych. W niektórych przypadkach różnice te można łatwo poprawić. Jeden zestaw danych może traktować imię i nazwisko jako pojedyncze pole, podczas gdy inny zestaw danych może używać osobnych pól dla imienia i nazwiska. Odpowiedź brzmi: zmień wszystkie zestawy danych, aby korzystały z jednego pola lub zmień je wszystkie, aby używały osobnych pól dla imienia i nazwiska. Niestety, wiele niedopasowań w zawartości danych jest trudniejszych do wykrycia. W rzeczywistości jest całkiem możliwe, że nie będziesz w stanie ich w ogóle zrozumieć. Zanim jednak się poddasz, rozważ następujące potencjalne rozwiązania problemu:

* Oblicz brakujące dane na podstawie innych danych, do których masz dostęp.

* Znajdź brakujące dane w innym zestawie danych.

* Połącz zestawy danych, aby utworzyć całość, która zapewnia spójne pola.

* Zbierz dodatkowe dane z różnych źródeł, aby uzupełnić brakujące dane.

* Przedefiniuj swoje pytanie, aby nie potrzebować już brakujących danych.

Oddzielanie użytecznych danych od innych danych

Niektóre organizacje są zdania, że nigdy nie mogą mieć zbyt dużej ilości danych, ale nadmiar danych staje się tak samo problemem, jak niewystarczającym. Aby skutecznie rozwiązywać problemy, AI wymaga tylko wystarczającej ilości danych. Zdefiniowanie pytania, na które chcesz udzielić zwięzłej i jasnej odpowiedzi, pomaga, podobnie jak użycie prawidłowego algorytmu (lub zestawu algorytmów). Oczywiście głównym problemem związanym z posiadaniem zbyt dużej ilości danych jest to, że znalezienie rozwiązania (po przejściu przez te wszystkie dodatkowe dane) trwa dłużej, a czasem wyniki są mylące, ponieważ nie widać lasu dla drzew.

W ramach tworzenia zestawu danych potrzebnego do analizy tworzona jest kopia oryginalnych danych, a nie modyfikowana. Zawsze zachowuj czyste, pierwotne dane, abyś mógł je później wykorzystać do innych analiz. Ponadto utworzenie odpowiednich danych wyjściowych do analizy może wymagać wielu prób, ponieważ może się okazać, że dane wyjściowe nie spełniają Twoich potrzeb. Chodzi o to, aby utworzyć zestaw danych, który zawiera tylko dane potrzebne do analizy, ale należy pamiętać, że dane mogą wymagać określonych rodzajów przycinania, aby zapewnić pożądany wynik.

Biorąc pod uwagę Pięć Mistruths w danych

Ludzie są przyzwyczajeni do przeglądania danych w wielu przypadkach: opinii. W rzeczywistości w niektórych przypadkach ludzie przekrzywiają dane do tego stopnia, że stają się bezużyteczne, nieprawdą. Komputer nie potrafi odróżnić prawdziwych od nieprawdziwych danych - widzi tylko dane. Jednym z problemów, które utrudniają, a nawet uniemożliwiają stworzenie sztucznej inteligencji, która tak naprawdę myśli jak człowiek, jest to, że ludzie mogą pracować z niepoprawnością, a komputery nie. Najlepsze, co możesz osiągnąć, to zobaczyć błędne dane jako wartości odstające, a następnie je odfiltrować, ale ta technika niekoniecznie rozwiązuje problem, ponieważ człowiek nadal używa danych i próbuje ustalić prawdę na podstawie nieścisłości, które są tam. Powszechną myślą o tworzeniu mniej zanieczyszczonych zestawów danych jest to, że zamiast pozwalać ludziom na wprowadzanie danych, zbieranie danych za pomocą czujników lub innych środków powinno być możliwe. Niestety czujniki i

inne metodologie wprowadzania mechanicznego odzwierciedlają cele ich ludzkich wynalazców i ograniczenia tego, co konkretna technologia jest w stanie wykryć. W związku z tym nawet dane pochodzące z komputera lub zmysłowa podlegają również generowaniu niepoprawności, które AI jest dość trudne do wykrycia i pokonania. W poniższych sekcjach wykorzystano wypadek samochodowy jako główny przykład do zilustrowania pięciu rodzajów nieścisłości, które mogą pojawić się w danych. Pojęcia, które wypadek próbuje przedstawić, nie zawsze pojawiają się w danych i mogą pojawiać się na różne sposoby niż omówione. Faktem jest, że zwykle przeglądając dane, musisz sobie z tym poradzić.

Popełnienie

Niepoprawności popełnienia to te, które odzwierciedlają całkowitą próbę zastąpienia prawdziwych informacji nieprawdziwymi. Na przykład, wypełniając raport z wypadku, ktoś może stwierdzić, że słońce chwilowo oślepiło ich, uniemożliwiając zobaczenie osoby, którą uderzyli. W rzeczywistości być może ta osoba była rozproszona przez coś innego lub tak naprawdę nie myślała o prowadzeniu samochodu (być może rozważając miły obiad). Jeśli nikt nie może obalić tej teorii, osoba ta może się liczyć z mniejszą karą. Chodzi jednak o to, że dane również zostałyby skażone. Skutkuje to tym, że teraz firma ubezpieczeniowa opierałaby składki na błędnych danych.

Chociaż wydaje się, że nie można całkowicie nie dopuścić do popełnienia morderstwa, często tak nie jest. Ludzie mówią „małe białe kłamstwa", aby ocalić innych od wstydu lub rozwiązać problem przy minimalnym wysiłku osobistym. Czasami nieprawda prowizji opiera się na błędnych informacjach lub przesłuchaniach. W rzeczywistości źródeł błędów popełnienia jest tak wiele, że naprawdę trudno jest wymyślić scenariusz, w którym ktoś mógłby ich całkowicie uniknąć. To powiedziawszy, niewierność popełnienia jest jednym z rodzajów niepoprawności, którego ktoś może uniknąć częściej niż innych.

Pominięcie

Niepoprawność pominięcia to te, w których człowiek mówi prawdę w każdym stwierdzonym fakcie, ale pomija ważny fakt, który zmieniłby postrzeganie incydentu jako całości. Myśląc ponownie o raporcie z wypadku, powiedz, że ktoś uderza jelenia, powodując znaczne uszkodzenie samochodu. Mówi zgodnie z prawdą, że droga była mokra; było już blisko zmierzchu, więc światło nie było tak dobre, jak mogłoby być; trochę spóźnił się z naciśnięciem hamulca; i jeleń po prostu wybiegł z zarośli na poboczu drogi. Wniosek byłby taki, że incydent jest po prostu wypadkiem. Jednak osoba pominęła ważny fakt. W tym czasie pisał SMS-y. Gdyby organy ścigania wiedziały o SMS-ach, zmieniłoby przyczynę wypadku na nieuważną jazdę. Kierowca może zostać ukarany grzywną, a ubezpieczyciel użyje innego powodu przy wprowadzaniu zdarzenia do bazy danych. Podobnie jak w przypadku nieścisłości prowizji, powstałe błędne dane zmieniłyby sposób, w jaki firma ubezpieczeniowa dostosowuje składki. Unikanie nieprawdy pomijania jest prawie niemożliwe. Tak, ktoś może celowo pominąć fakty w raporcie, ale równie prawdopodobne jest, że ktoś po prostu zapomni podać wszystkie fakty. W końcu większość ludzi jest bardzo zaniepokojona po wypadku, więc łatwo stracić koncentrację i zgłosić tylko te prawdy, które wywarły największe wrażenie. Nawet jeśli później osoba zapamięta dodatkowe szczegóły i je zgłosi, jest mało prawdopodobne, aby baza danych zawierała pełny zestaw prawd.

Perspektywiczny

Niepoprawność perspektywy pojawia się, gdy wiele stron ogląda incydent z wielu punktów widokowych. Na przykład przy rozważaniu wypadku z udziałem uderzonego pieszego, osoby prowadzącej samochód, osoby potrąconej przez samochód oraz osoby postronnej, która była świadkiem tego wydarzenia, miałyby różne perspektywy. Oficer sporządzający raporty od każdej

osoby, co zrozumiałe, uzyskałby różne fakty z każdej z nich, nawet zakładając, że każda osoba mówi prawdę taką, jaką zna. W rzeczywistości doświadczenie pokazuje, że prawie zawsze tak jest, a to, co oficer przedstawia jako raport, stanowi podstawę tego, co każdy z zaangażowanych, wzbogacony osobistym doświadczeniem. Innymi słowy, raport będzie zbliżony do prawdy, ale niewystarczająco bliski dla sztucznej inteligencji. W przypadku perspektywy ważne jest, aby wziąć pod uwagę punkt obserwacyjny. Kierowca samochodu widzi deskę rozdzielczą i zna stan samochodu w momencie wypadku. Jest to informacja, której brakuje pozostałym dwóm stronom. Podobnie osoba, która została potrącona przez samochód, ma najlepszy punkt obserwacyjny do zobaczenia wyrazu twarzy (intencji) kierowcy. Obserwujący może być w najlepszej pozycji, aby sprawdzić, czy kierowca próbował się zatrzymać, i ocenić takie kwestie, jak to, czy kierowca próbował skręcić. Każda ze stron będzie musiała sporządzić raport na podstawie widocznych danych bez korzyści z ukrytych danych.

Perspektywa jest być może najniebezpieczniejszą z niepoprawności, ponieważ każdy, kto spróbuje wydobyć prawdę w tym scenariuszu, w najlepszym razie skończy ze średnią z różnych historii, które nigdy nie będą w pełni poprawne. Człowiek oglądający informacje może polegać na intuicji i instynkcie, aby potencjalnie uzyskać lepsze przybliżenie prawdy, ale sztuczna inteligencja zawsze użyje tylko średniej, co oznacza, że sztuczna inteligencja zawsze ma znaczną wadę. Niestety, unikanie nieścisłości perspektywy jest niemożliwe, ponieważ bez względu na to, ilu świadków masz na to wydarzenie, najlepsze, co możesz mieć nadzieję osiągnąć, to przybliżenie prawdy, a nie prawdy faktycznej.

Jest też inny rodzaj nieufności do rozważenia i jest to jeden z nich perspektywiczny. Pomyśl o tym scenariuszu: w 1927 r. jesteś osobą niesłyszącą. Co tydzień chodzisz do teatru, żeby obejrzeć film niemy, i przez godzinę lub dłużej czujesz się jak wszyscy inni. Możesz oglądać film w taki sam sposób, jak wszyscy inni; nie ma różnic. W październiku tego roku zobaczysz znak, że teatr modernizuje się, aby obsługiwać system dźwiękowy, aby mógł wyświetlać talie - filmy ze ścieżką dźwiękową. Znak mówi, że jest to najlepsza rzecz na świecie i prawie wszyscy wydają się zgadzać, z wyjątkiem ciebie, osoby niesłyszącej, która ma teraz poczuć się jak obywatel drugiej kategorii, różni się od wszystkich innych, a nawet jest prawie wykluczona z teatru . W oczach osoby niesłyszącej znak ten jest niepoprawnością; dodanie systemu dźwiękowego jest najgorszą możliwą rzeczą, a nie najlepszą możliwą. Chodzi o to, że to, co wydaje się ogólnie prawdą, wcale nie jest prawdą dla wszystkich. Idea ogólnej prawdy - która jest prawdziwa dla wszystkich - jest mitem. Nie istnieje.

Stronniczość

Nieufność wobec uprzedzeń występuje, gdy ktoś jest w stanie zobaczyć prawdę, ale z powodu osobistych obaw lub przekonań nie jest w stanie jej zobaczyć. Na przykład, myśląc o wypadku, kierowca może skupić uwagę tak całkowicie na środku drogi, że jeleń na skraju drogi staje się niewidoczny. W związku z tym kierowca nie ma czasu na reakcję, gdy jeleń nagle postanawia wybiec na środek drogi, próbując przejść. Problem z uprzedzeniami polega na tym, że kategoryzacja może być niezwykle trudna. Na przykład kierowca, który nie widzi jelenia, może mieć prawdziwy wypadek, co oznacza, że jeleń został ukryty przez krzaki. Jednak kierowca może być również winny nieuważnej jazdy z powodu nieprawidłowego ustawienia ostrości. Kierowca może również doświadczyć chwilowego rozproszenia uwagi. Krótko mówiąc, fakt, że kierowca nie widział jelenia, nie jest pytaniem; zamiast tego zależy, dlaczego kierowca nie widział jelenia. W wielu przypadkach potwierdzenie źródła błędu systematycznego staje się ważne podczas tworzenia algorytmu zaprojektowanego w celu uniknięcia źródła błędu wstępnego. Teoretycznie unikanie nieścisłości stronniczości jest zawsze możliwe. W rzeczywistości jednak wszyscy ludzie mają różne typy uprzedzeń, które zawsze skutkują nieścisłościami, które wypaczają zestawy danych. Nakłonienie kogoś, aby faktycznie spojrzał, a następnie zobaczył coś - aby zarejestrował się w mózgu tej osoby - jest trudnym zadaniem. Ludzie polegają na filtrach, aby

uniknąć przeciążenia informacji, a filtry te są również źródłem stronniczości, ponieważ uniemożliwiają ludziom widzenie rzeczy.

Ramy Odniesienia

Z pięciu niewiadomych układ odniesienia nie musi być wynikiem jakiegokolwiek błędu, ale zrozumienia. Niepoprawność ram odniesienia ma miejsce, gdy jedna ze stron opisuje coś, na przykład zdarzenie takie jak wypadek, a ponieważ druga strona nie ma doświadczenia w tym wydarzeniu, szczegóły stają się mętne lub całkowicie niezrozumiane. Istnieje wiele procedur komediowych, które opierają się na błędach ramy odniesienia. Jednym ze znanych przykładów jest Abbott i Costello, Who's On First ?. Zrozumienie, co mówi druga osoba, może być niemożliwe, gdy pierwszej osobie brakuje wiedzy empirycznej - ramy odniesienia. Inny przykład nieścisłości z ramami odniesienia występuje, gdy jedna ze stron nie jest w stanie zrozumieć drugiej. Na przykład marynarz doświadcza burzy na morzu. Być może jest to monsun, ale załóżmy przez chwilę, że burza jest znaczna - być może zagraża życiu. Nawet przy użyciu filmów, wywiadów i symulatora doświadczenie przebywania na morzu w burzy zagrażającej życiu byłoby niemożliwe do przekazania komuś, kto nie doświadczył takiej burzy z pierwszej ręki; ta osoba nie ma układu odniesienia.

Najlepszym sposobem na uniknięcie nieścisłości związanych z ramami odniesienia jest zapewnienie wszystkim zainteresowanym stronom możliwości opracowania podobnych ram odniesienia. Aby zrealizować to zadanie, różne strony wymagają podobnej wiedzy doświadczalnej, aby zapewnić dokładne przesyłanie danych od jednej osoby do drugiej. Jednak podczas pracy z zestawem danych, który jest koniecznie rejestrowany, dane statyczne, nadal występują błędy ramy odniesienia, gdy potencjalny widz nie ma wymaganej wiedzy doświadczalnej. AI zawsze będzie mieć problemy z ramą odniesienia, ponieważ AI niekoniecznie nie ma możliwości stworzenia doświadczenia. Baza danych nabytej wiedzy to nie to samo. Bank danych zawierałby fakty, ale doświadczenie opiera się nie tylko na faktach, ale także na wnioskach, których nie można powielić obecnej technologii.

Definiowanie limitów pozyskiwania danych

Może się wydawać, że wszyscy zdobywają twoje dane bez zastanowienia i bez powodu, i masz rację. W rzeczywistości organizacje gromadzą, kategoryzują i przechowują dane wszystkich osób - pozornie bez celu lub zamiaru. Według Data Never Sleeps świat zbiera dane w tempie 2,5 kwintillionów bajtów dziennie. Te codzienne dane są dostępne w różnych formach, co potwierdzają poniższe przykłady:

* Google przeprowadza 3 607 080 wyszukiwań.

* Użytkownicy Twittera wysyłają 456 000 tweetów.

* Użytkownicy YouTube oglądają 4 146 600 filmów.

* Skrzynki odbiorcze odbierają 103 447 529 e-maili ze spamem.

* Kanał pogody odbiera 18 055 555,56 wniosków o pogodę.

* GIPHY obsługuje 694 444 GIF-y.

Akwizycja danych stała się narkotykiem dla organizacji na całym świecie, a niektórzy uważają, że organizacja, która zbiera najwięcej, jakoś wygrywa nagrodę. Jednak samo gromadzenie danych niczego nie osiąga. Książka „Autostopowicz po galaktyce" Douglasa Adamsa wyraźnie ilustruje ten problem. W tej książce rasa superkultur buduje ogromny komputer do obliczenia znaczenia „życia, wszechświata i wszystkiego". Odpowiedź 42 tak naprawdę niczego nie rozwiązuje, więc niektóre stworzenia narzekają, że zbieranie, kategoryzacja i analiza wszystkich danych użytych w odpowiedzi nie dały użytecznego

rezultatu. Komputer, nie tylko rozumiejący, mówi ludziom, że odpowiedź jest prawidłowa, ale muszą znać pytanie, aby odpowiedź miała sens. Pozyskiwanie danych może odbywać się w nieograniczonych ilościach, ale znalezienie odpowiednich pytań może być zniechęcające, jeśli nie niemożliwe. Głównym problemem, który musi rozwiązać każda organizacja w zakresie pozyskiwania danych, jest to, jakie pytania należy zadać i dlaczego są one ważne. Dostosowywanie akwizycji danych do odpowiedzi na pytania, na które potrzebujesz odpowiedzi. Na przykład, jeśli prowadzisz sklep w mieście, możesz potrzebować takich pytań, na które odpowiedziałeś:

* Ile osób chodzi codziennie przed sklepem?

* Ilu z tych ludzi przestaje patrzeć w okno?

* Jak długo wyglądają?

* O której godzinie patrzą?

* Czy niektóre wyświetlacze zwykle dają lepsze wyniki?

* Który z tych wyświetlaczy powoduje, że ludzie wchodzą do sklepu aby robić zakupy?

Lista może trwać dalej, ale pomysł polega na tym, że stworzenie listy pytań, które odpowiadają konkretnym potrzebom biznesowym, jest niezbędne. Po utworzeniu listy musisz sprawdzić, czy każde z pytań jest rzeczywiście ważne - to znaczy zaspokoić potrzebę - a następnie ustalić, jakiego rodzaju informacji potrzebujesz, aby odpowiedzieć na pytanie. Oczywiście próba zebrania wszystkich tych danych ręcznie byłaby niemożliwa, i tu właśnie pojawia się automatyzacja. Pozornie automatyzacja zapewniłaby niezawodne, powtarzalne i spójne wprowadzanie danych. Jednak wiele czynników w automatyzacji akwizycji danych może dawać dane, które nie są szczególnie przydatne. Weźmy na przykład następujące problemy:

* Czujniki mogą gromadzić tylko te dane, które mają gromadzić, więc możesz przegapić dane, gdy używane czujniki nie są przeznaczone do cel, powodu.

* Ludzie tworzą błędne dane na różne sposoby, co oznacza, że otrzymywane dane mogą być fałszywe.

* Dane mogą ulec zniekształceniu, gdy istnieją warunki ich gromadzenia niepoprawnie zdefiniowane.

* Niepoprawna interpretacja danych oznacza, że wyniki również będą błędne.

* Przekształcanie pytania ze świata rzeczywistego w algorytm komputera rozumie, że jest to proces podatny na błędy.

Należy wziąć pod uwagę wiele innych kwestii (wystarczających do wypełnienia książki). Łącząc źle zebrane, źle sformułowane dane z algorytmami, które w rzeczywistości nie odpowiadają na pytania, uzyskuje się dane wyjściowe, które mogą faktycznie prowadzić Twoją firmę w złym kierunku, dlatego AI jest często obwiniane za niespójne lub nierzetelne wyniki. Zadanie właściwego pytania, uzyskanie prawidłowych danych, właściwe przetwarzanie, a następnie poprawna analiza danych są wymagane, aby pozyskiwanie danych było narzędziem, na którym można polegać.