DIGITAL SUSTAINABILITY FORUM

Sztuczna inteligencja jako element zrównoważonego rozwoju gospodarczo – społecznego

Raport podsumowujący drugie spotkanie w ramach projektu Digital Sustainability Forum

Warszawa 2019

Partnerzy Microsoft digitalpoland CENTRUM CYFROWE

DIGITAL SUSTAINABILITY FORUM

Roundtable #2 - Data is the new oil

Cele spotkania

Tytuł drugiego roundtable w ramach projektu Digital Sustainability Forum (DSF) dotyczył ewolucji gospodarki w stronę jej nowego oblicza, opierającego się na wiedzy i danych. Większość rewolucyjnych modeli biznesowych w nowej gospodarce pozwala zbudować przewagę konkurencyjną właśnie dzięki lepszemu wykorzystaniu danych m.in. w decyzjach biznesowych. Nie inaczej jest w przypadku rozwiązań wykorzystujących sztuczną inteligencję (AI), w których wynik końcowy, w postaci nauczonego algorytmu, zależy w bardzo dużym stopniu od użytych danych podczas procesu uczenia algorytmu (proces ten potocznie nazywa się „treningiem algorytmu").

W kolejnej sesji DSF skupiliśmy się na wypracowaniu spójnego punktu widzenia na temat znaczenia danych w nowoczesnej gospodarce, ich dostępności oraz zasad wykorzystywania. Naszym celem było przygotowanie rekomendacji dla budowy zaleceń, kodeksów branżowych i propozycji regulacji ułatwiających wykorzystywanie danych w gospodarce. Wierzymy, że wpłynie to na rozwój i wykorzystywanie rozwiązań AI w Polsce.

W dyskusji skupiliśmy się na:

- ocenie zasobów danych w polskiej gospodarce,
- określeniu barier i katalizatorów wpływających na gromadzenie danych oraz zdefiniowaniu rekomendacji w tym zakresie,
- wskazaniu modeli biznesowych w gospodarce opartej o danych, które wesprą dzielenie się danymi,
- określeniu zasad współpracy, w tym dostępu, współdzielenia i wykorzystywania danych pomiędzy partnerami ekosystemu AI,
- zdefiniowaniu portfolio modeli zarządzania danymi – data governance i rekomendacji ich obszarów stosowania,
- wypracowaniu rekomendacji w zakresie zasad otwierania danych publicznych,

- wskazaniu sektorów gospodarki kluczowych z punktu widzenia tworzenia i wykorzystania zbiorów danych,
- określeniu niezbędnych ram regulacji w celu zachowania równowagi pomiędzy działaniami organizującymi i porządkującymi, a tworzeniem zbyt daleko idących ograniczeń,

Poniżej przedstawiamy podsumowanie dyskusji i efektów pracy analitycznej pogłębiające i porządkujące kwestie wykorzystania danych do budowy i wykorzystania rozwiązań AI.

Spis Treści

"O" – Objective - Cel

Celem była ocena struktury i zasobów danych w gospodarce oraz przygotowanie rekomendacji dla budowy zaleceń, kodeksów branżowych i propozycji regulacji ułatwiających wykorzystywanie danych w gospodarce.

"S" - Status

Definicja i klasyfikacja danych

Tradycyjne definicje danych skupiają się głównie na danych przygotowanych i wykorzystywanych przez człowieka, określając je jako fakty i zbiory statystyczne wykorzystywane do analiz (*za Oxford Dictionary*).

Z perspektywy systemów komputerowych ta **definicja wymaga poszerzenia i odnosi się najczęściej do zapisu wszystkich informacji w formie cyfrowej do wykorzystywania przez komputery**.

Z perspektywy rozwoju analityki, Big Data i sztucznej inteligencji (AI) **należy poświęcić szczególną uwagę komputerowym danym tworzonym w dużej ilości nie tylko przez ludzi**, ale generowanych przez różnego rodzaju systemy i urządzenia. Dane mogą bowiem pochodzić z kamer, mikrofonów, różnorodnych sensorów i mierników zapisujących stan urządzenia, czy też z wewnętrznych zapisów komputerowych (logów). Sensory przy tym szczególnie często zapisują stan z dużą częstotliwością próbując oddać przebieg zjawisk w czasie rzeczywistym.

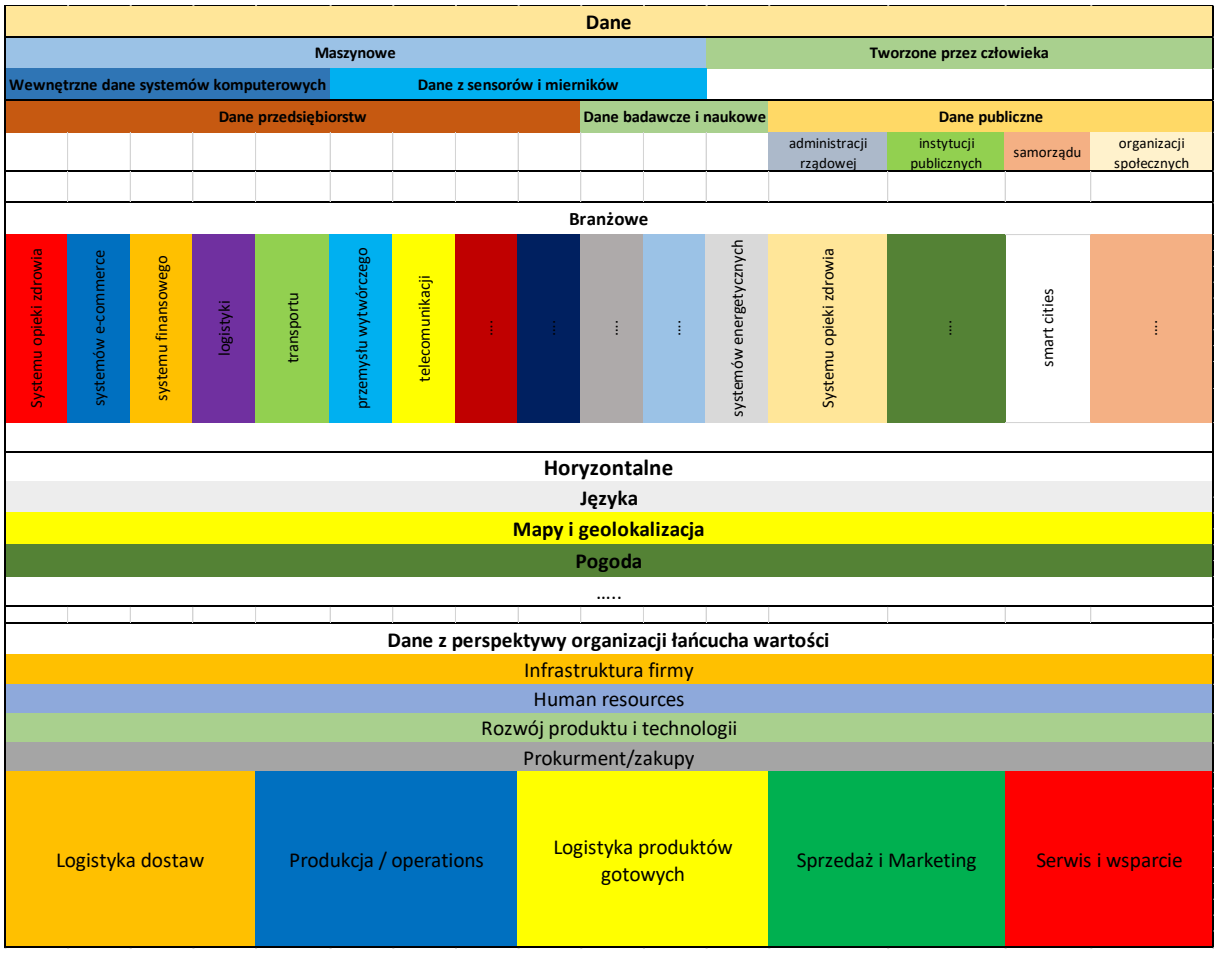
Do **danych maszynowych** należy dodać zapisy **tworzone przez ludzi** w formie baz danych i rejestrów.

Dane też można dzielić ze względu na źródła pochodzenia i główne obszary wykorzystywania, np.: **dane publiczne** (dane z administracji rządowej, dane z jednostek samorządu terytorialnego, dane z organizacji społecznych), **dane badawcze** i naukowe, **dane przedsiębiorstw**.

Innym podziałem jest podział na dane o charakterze **horyzontalnym** i **branżowym**. Dane o charakterze **horyzontalnym** to dane, które są wykorzystywane przez szerokie spektrum podmiotów z różnych branż. Do takich danych należą bazy języka pisanego i mówionego, dane geograficzne i geolokalizacyjne, dane o pogodzie, jakości powietrza, zdjęcia otoczenia. Dane o charakterze **branżowym** to dane tworzone i wykorzystywane w poszczególnych dziedzinach gospodarki i życia społecznego np. dane systemu ochrony zdrowia (w tym dane medyczne pacjentów), dane z systemów energetycznych i przesyłowych, dane z sektora telekomunikacyjnego, dane z sektora finansowego, dane z portali informacyjnych, systemów e-commerce. **Na szczególną uwagę**, oprócz danych przedsiębiorstw pogrupowanych w branżach, **zasługują dane z systemów służących do zarządzania miastami**, w takich obszarach jak: transport (tzw. systemy ITS), bezpieczeństwo publiczne, zarządzanie infrastrukturą, ochrona środowiskiem czy jakość życia mieszkańców. Wykorzystanie ww. systemów jest jednym z kluczowych elementów powstania tzw. inteligentnych miast (ang. *smart city*).

Na dane przedsiębiorstw i branż można również spojrzeć z perspektywy **odwzorowania łańcucha tworzenia wartości**.

Poniższy rysunek może przedstawić różnorodność klasyfikacji danych. Intencją rysunku nie jest pokazanie wszystkich wymiarów i pełnego podziału w ramach poszczególnych przekrojów. Pomaga on jednak zrozumieć złożoność problemu i porządkuje wybrane kwestie.



Rysunek 1. Różnorodność klasyfikacji danych. Źródło: Opracowanie własne.

Opracowany na zlecenie Ministerstwa Przedsiębiorczości i Technologii raport pt.: „Zasoby *danych w polskiej gospodarce, które mogą służyć zwiększaniu konkurencyjności polskich firm*" wnosi istotny wkład do dyskusji na temat danych możliwych do wykorzystania w gospodarce.

Raport skupia się jednak na wybranych obszarach i przyjmuje założenia wykluczające z zakresu raportu istotne dla AI obszary np. dane przedsiębiorstw oraz dane tworzone przez systemy w samorządach w obszarze tzw. inteligentnych miast.

W konsekwencji **raport nie analizuje danych z perspektywy branżowej, nie analizuje również danych z perspektywy łańcucha tworzenia wartości, skupiając się głównie na obszarze danych tworzonych przez człowieka**. W obszarach, w których raport dyskutuje kwestie związane ze sztuczną inteligencją i uczeniem maszynowym, **odnosi się w zasadzie do zagranicznych przykładów i opisu zagranicznych źródeł danych**. Próba odniesienia zakresu raportu do zakresu prezentowanego na rysunku powyżej pozwala stworzyć poniżej przybliżony obraz.

Dane														
Maszynowe										**Tworzone przez człowieka**				
Wewnętrzne dane systemów komputerowych			**Dane z sensorów i mierników**											
Dane przedsiębiorstw							**Dane badawcze i naukowe**			**Dane publiczne**				
										administracji rządowej	instytucji publicznych	samorządu	organizacji społecznych	

Branżowe														
Systemu opieki zdrowia	systemów e-commerce	systemu finansowego	logistyki	transportu	przemysłu wytwórczego	telecomunikacji	....	....	....	....	systemów energetycznych	Systemu opieki zdrowia	....	smart cities	....

Horyzontalne
Języka
Mapy i geolokalizacja
Pogoda
....

Dane z perspektywy organizacji łańcucha wartości				
Infrastruktura firmy				
Human resources				
Rozwój produktu i technologii				
Prokurment				
Logistyka dostaw	Produkcja / operations	Logistyka produktów gotowych	Sprzedaż i Marketing	Serwis i wsparcie

Rysunek 2. Zakres raportu wykonanego dla MPiT – zaznaczenie kolorem zielonym. Źródło: Opracowanie własne.

Rola danych w budowie i zastosowaniach rozwiązań sztucznej inteligencji

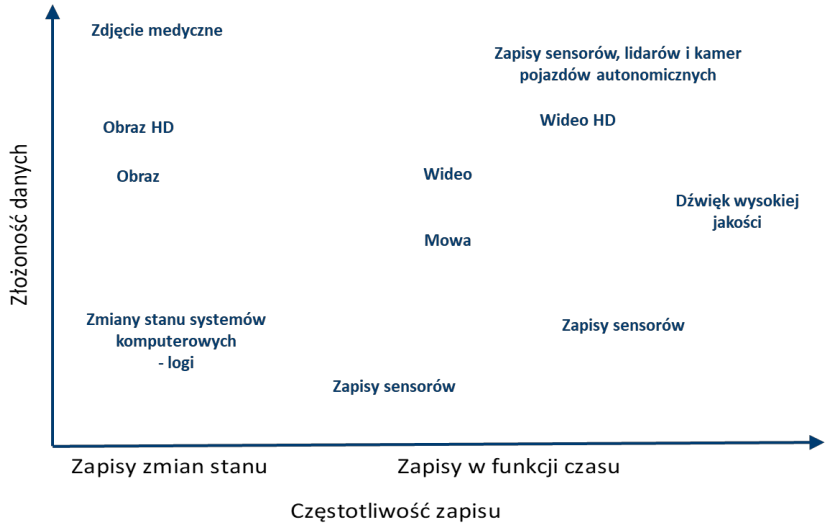
Dominującym jest pogląd, że rozwiązania AI zależą wprost od jakości i rozmiarów dostępnych danych.

Wielkość zbiorów danych zależy od złożoności danych, zależności od czasu oraz od długości sekwencji danych z punktu widzenia tworzenia rozwiązań wykorzystujących uczenie maszynowe.

Złożoność danych to ilość i różnorodność danych potrzebnych do opisu stanu w jednej chwili czasu i może przyjmować wielkości od kliku bitów informacji np. w zapisach stanu sensora on/off, do kilkuset MB lub nawet GB w przypadku zapisów obrazów o dużej wielkości i dużej rozdzielczości.

Zależność od czasu wynika z decyzji czy twórca rozwiązania zapisze zmiany stanu wybranych parametrów, pomijając zapis pomiędzy zmianami np. zapisy transakcji, logów w systemach komputerowych, czy twórca zapisze pełny przebieg czasowy wybranych parametrów. Pełny przebieg czasowy można zapisywać z różną częstotliwością.

Częstotliwość zapisu wynika z wyboru jakości odwzorowania zmian wybranych wielkości w czasie i może przyjmować różne wielkości np. niewielkie w zapisach geologicznych, meteorologicznych, zapisach zmian w parametrów do kilkudziesięciu zapisów na sekundę w przypadku filmów HD, niemal 200 tysięcy zapisów na sekundę w przypadku zapisu dźwięku wysokiej jakości, a specjalistyczne kamery i inne sensory mogą zapisywać dane jeszcze częściej.



Rysunek 3. Generowanie danych. Źródło: Warsztat transformacji cyfrowej; Tomasz Klekowski; Kwiecień 2019

Kolejną **istotną kwestią jest długość sekwencji potrzebnej do trenowania rozwiązań sztucznej inteligencji**. Jeff Dean Google Senior Fellow i szef projektu Google Brain wskazał w wywiadzie z października 2017, że aby myśleć o rozwiązaniach głębokiego uczenia należy mieć do dyspozycji sekwencje danych o długości 100 000 elementów.

Podobny rząd wielkości można spotkać analizując dyskusje dotyczące uczenia maszynowego w serwisie Research Gate (www.reserchgate.net) gdzie omawiany w czerwcu 2018r model rozpoznawania obrazu dla celów sterowania pojazdem, przygotowany w firmie Jaguar Land Rover, od liczby 80 000 elementów zaczynał dawać pozytywne rezultaty, które polepszyły się dla 140 000 elementów, a autorzy oczekiwali, że docelowy trening algorytmów będzie oparty na milionie elementów.

Należy zaznaczyć, że pojawia się znacząca rozbieżność pomiędzy wielkością zbiorów używanych w działalności akademickiej, gdzie zbiory są mniejsze, a w działalności komercyjnej, gdzie wskazywane są większe liczby. Informacje uzyskane od polskich startupów wskazują, że pracują one często w oparciu o zbiory danych rzędu 10 000 elementów.

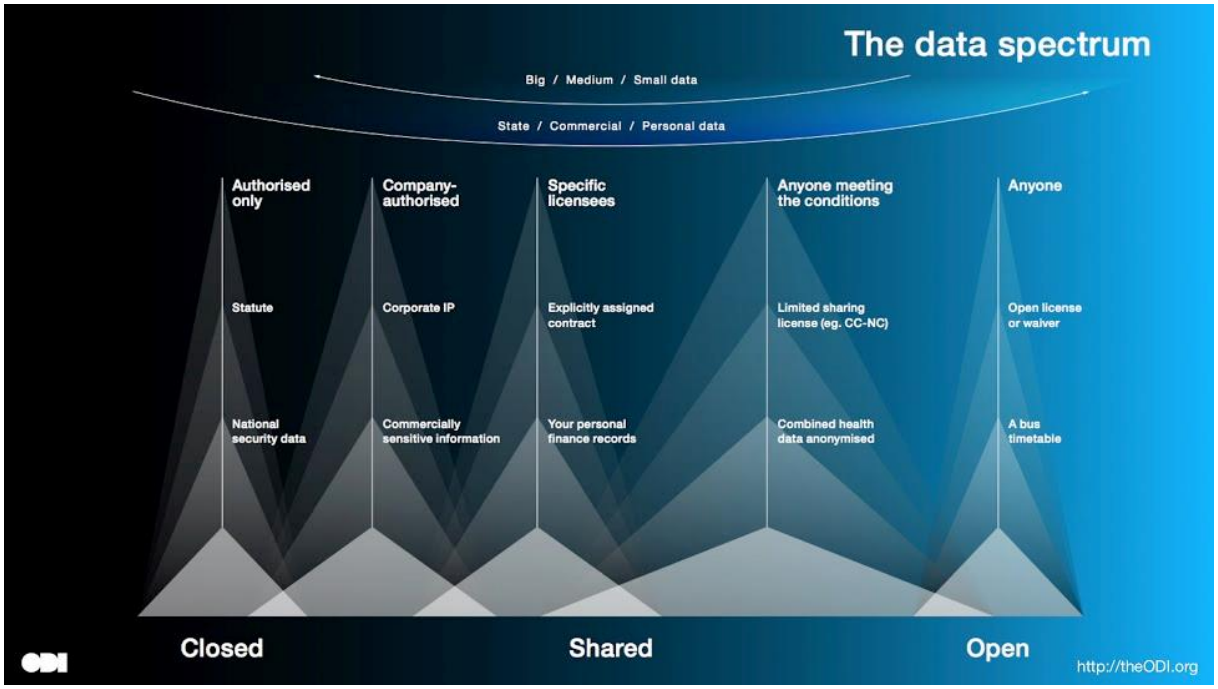
Ilość danych wymaganych do stworzenia i wytrenowani rozwiązania sztucznej inteligencji zależy również od przyjętego sposobu trenowania modelu.

Problem niedostatecznej ilości danych można często rozwiązać poprzez komputerową syntezę danych. Dane syntetyczne to dane generowane komputerowo, które naśladują rzeczywiste dane np. transformacje kolorystyczne zdjęć, zmiany perspektywy zdjęć, symulacje z silników gier komputerowych.

Drugim po ilości decydującym aspektem jest jakość danych, oddająca różnorodność i reprezentatywność cech populacji w próbce. Brak wystarczającej jakości i reprezentatywności danych może skutkować, skrzywieniem modelu (ang. *bias*). Skutki skrzywienia modelu mogą się rozciągać wzdłuż szerokiego spektrum: od jego drobnej niedokładności, co może skutkować pogorszoną jakością produktu, po bardzo istotne problemy powstające w systemach przetwarzających wrażliwe dane osobowe, gdzie skrzywienie może skutkować dyskryminacją, zagrożeniem bezpieczeństwa ludzi i poważnymi skutkami prawnymi.

Klasyfikacja danych z punktu widzenia dostępu

Open Data Institute (ODI) z Wielkiej Brytanii w raporcie "*The role of data in AI business models*" z kwietnia 2018 roku prezentuje następujący model klasyfikacji danych z punktu widzenia dostępu:



Rysunek 4. Klasyfikacja danych. Źródło: Open Data Institute

Z perspektyw firmy to kontroluje ona dane wewnętrzne/zamknięte i firmom jest je najłatwiej wykorzystać. Łatwo jest również wykorzystać dane otwarte, jednak mogą się w tym obszarze pojawić wyzwania związane z ich jakością, opisem ułatwiającym naukę algorytmu i przygotowaniem do wykorzystania w określonym rozwiązaniu. Największej uwagi wymagają dane współdzielone, które mogą być podstawą do agregacji i korelacji z innymi zbiorami danych i z racji potencjału budowy dużych i różnorodnych zbiorów danych mogą stanowić podstawę do tworzenia nowych rozwiązań. Aby mogło się tak stać należy zdefiniować, opisać i przygotować odpowiednie modele tworzenia takich zasobów, ich zarządzania oraz udostępniania i dzielenia rezultatów pracy w oparciu o te dane.

Źródła danych wykorzystywane przez polskie startupy

W celu uzyskania bardziej precyzyjnego obrazu i oceny stopnia dostępności danych w Polsce należy spojrzeć głębiej i przeanalizować poszczególne obszary zastosowania sztucznej inteligencji.

	Obszary	Przykłady
Zastosowania AI	Systemy analityczne i predykcyjne	Predictive maintanance Systemy rekomendacji
	Systemy optymalizacji i planowania	Logistyka, planowanie produkcji
	NLP – Natural Language Processing	Automatyczne tłumaczenie, analiza tekstów, tworzenie treści, chatbooty
	Rozpoznawanie i synteza mowy	Chatboty, asystenci
	Robotyka	Robotyka i Robotic Process Automation
	Przetwarzanie obrazu,	Rozpoznawanie obrazów, widzenie maszynowe

Rysunek 5. Zastosowania AI. Źródło: Warsztat transformacji cyfrowej; Tomasz Klekowski; Kwiecień 2019.

Na potrzeby warsztatu zostało przeprowadzonych kilka wywiadów z firmami, reprezentującymi poszczególne kategorie rozwiązań. Zidentyfikowano podstawowe źródła danych wykorzystywane przez firmy tworzące rozwiązania AI w Polsce:

1. Systemy analityczne, predykcyjne i optymalizacyjne - głównie wykorzystują dane z przedsiębiorstw. Dane pochodzące z baz danych transakcji, logów komputerowych i sensorów;
2. Systemy NLP i chatboty - wykorzystują teksty z komunikatorów internetowych oraz nagrania i zapisy rozmów z centrów telefonicznych przedsiębiorstw (ang. *call centers*);
3. Systemy rozpoznawania obrazów oraz systemy analizy wideo i widzenia maszynowego - wykorzystują zdjęcia i zapisy wideo archiwizowane w monitoringu, zdjęcia dostępne w internecie oraz w modelu DaaS;
4. Systemy robotyczne, w tym pojazdy autonomiczne, oraz systemy RPA – dane tworzone są podczas testów i wdrożeń.

Systemy analityczne, predykcyjne i optymalizacyjne

W tej grupie znajdują się rozwiązania wielu branż (finansowej, e-commerce, telekomunikacyjnej, produkcyjnej, cybersecurity, systemów CRM) oraz większość obszarów inteligentnych miast. Dane pochodzą z komputerowych systemów przedsiębiorstw, zapisów transakcji, logów komputerowych, baz danych oraz sensorów. Wykorzystanie sensorów związane jest z odwzorowaniem fizycznych procesów.

Systemy analityczne i rekomendacyjne są drugim i trzecim obszarem z perspektywy powszechności zastosowań rozwiązań AI wg raportu „Map of the Polish AI" opublikowanego przez Fundację Digital Poland, ze wskazaniami 55% i 52%.

Dane z logów i systemów transakcyjnych, czy też bilingowych są podstawą budowania rozwiązań np. przez takie firmy jak np. Allegro, Synerise, Brand 24, DeepSense.

Dane przedsiębiorstw nie są udostępniane i są traktowane jako niematerialna własność firmy, natomiast dane z systemów inteligentnych miast w niektórych obszarach mogą być udostępniane do dalszego wykorzystania przez miasta. W Polsce prym w udostępnianiu danych z systemów miejskich wiedzie Warszawa oraz Gdańsk.

W przypadku danych pochodzących z sensorów rozmiary zbiorów danych zależą od rodzajów sensorów i częstotliwości zapisu. W sytuacjach kiedy występuje zapis z dużą częstotliwością powstają olbrzymie zbiory danych.

W testach inteligentnej infrastruktury drogowej wspierającej pojazdy autonomiczne przeprowadzonych przez firmę T-Systems w Berlinie w ramach projektu DIGINET-PS każdy sensor generuje ponad 300MB danych dziennie. Stworzenie takiej infrastruktury dla całego miasta wielkości Berlina wymagałoby użycia 150 000 sensorów i generowałoby **25 do 30 TB danych dziennie**. (Źródło: 5G Pushes Autonomous Driving, Joerg Tischler; T-Systems, Kwiecień 2019)

Projektowane dla wsparcia takich systemów centra danych budowane są z myślą o przetwarzaniu PB (10^{15}) danych i przechowywaniu EB (10^{18}) danych.

Barierą rozwoju tego obszaru w Polsce jest niski stopień świadomości większości polskich form w obszarze analityki danych oraz brak strategii i planu wykorzystania rozwiązań AI, które nie definiują obszarów wykorzystania tych technologii, a co za tym idzie nie definiują rodzaju, zakresu i formatów danych dla planowanych rozwiązań AI. Wymieniona wyżej firma T-Systems nie prowadzi w Polsce podobnych testów pomimo obecności na polskim rynku w ramach grupy T-Mobile.

Systemy NLP i chatboty

W obszarze tym mamy do czynienia z rozumieniem języka i przetwarzaniem mowy na tekst i tekstu na mowę. Te obszary należy traktować rozłącznie.

Obszary związane z analizą języka i jego rozumieniem są wykorzystywane w tłumaczeniu maszynowym oraz w chatbotach tekstowych. Chatboty głosowe wykorzystują dodatkowo moduły przetwarzające mowę na tekst i tekst na mowę.

Polskie firmy pracujące w tym obszarze w większości tworzą rozwiązania związane z rozumieniem języka i dopasowywaniem odpowiedzi, podczas gdy przetwarzanie mowy na tekst i tekstu na mowę jest najczęściej oparte o wykorzystanie narzędzi firmy Google

(Dialogflow oraz Cloud Speech-to-Text API), zapowiadany jest również podobny produkt opracowany przez firmę Amazon na potrzeby polskiej wersji asystenta Alexa.
Wobec nieodpłatnego (lub bardzo niskiej ceny) oferowania swoich usług przez duże firmy technologiczne niejasna jest przyszłość polskich firm pracujących w obszarze speech to text np. Stanusch Technologies, Senti.one, Inteliwise.
Osobną kwestią jest sprawa decyzji czy tworzenie narzędzi speech to text oraz text to speech dla języka polskiego powinna być traktowana wyłącznie komercyjnie, czy państwo polskie powinno stworzyć i wspierać otwarte narzędzia w tym obszarze.
Z perspektywy źródeł pochodzenia danych widać, że firmy wykorzystują teksty z komunikatorów internetowych oraz nagrania i zapisy rozmów z centrów telefonicznych przedsiębiorstw (ang. *call center*). Najczęściej chatboty są wykorzystywane w branżach, gdzie funkcjonują prawne regulacje dotyczące zawierania umów na odległość i związana z tym rejestracja rozmów oraz tam gdzie kontakt odbywa się za pomocą komunikatorów.
Raport „Map of the Polish AI" wskazuje, że 45% firm AI w Polsce zajmuje się przetwarzaniem języka, a 17% zajmuje się rozpoznawaniem mowy.

Systemy rozpoznawania obrazów oraz systemy analizy wideo i widzenia maszynowego

Rozpoznawanie obrazów jest jednym z najbardziej popularnych obszarów zastosowań AI. Najwięcej firm, w raporcie „Map of the Polish AI", 62% wskazuje, że tworzy rozwiązania w tym obszarze. Do tej grupy można na przykład zaliczyć systemy rozpoznawanie defektów na zdjęciach produktów, systemy analizy zdjęć medycznych, systemy rozpoznawania twarzy, systemy rozpoznawania stopnia zmęczenia kierowcy, systemy analizy grup klientów sieci handlowych, systemy bezpieczeństwa miejsc publicznych.
Z perspektywy danych rozpoznawanie obrazów metodą uczenia z nadzorem (ang. *supervised learning*) wymaga treningu modelu w oparciu o zbiór opisanych zdjęć. Jakość, różnorodność i reprezentatywność tych zdjęć będzie wpływała na jakość działania systemu. Opisywanie zdjęć jest czasochłonnym zadaniem i dodatkowo zależy od celu postawionego przed modelem. Dlatego w tej grupie rozwiązań **często tworzy się obrazy/zdjęcia dla konkretnych rozwiązań** i opisuje się je na potrzeby treningu. Dodatkowo można zwiększać liczbę zdjęć wykorzystywanych do treningu modelu stosując różne techniki np. zmiany kolorów, perspektywy, naświetlenie etc.
Rozmiary zbiorów danych różnią się i często można rozpocząć wykorzystywanie rozwiązań tworzonych w oparciu o zbiory kilku tysięcy obrazów, ale komercyjne zastosowania

wymagają zbiorów większych, jak w omawianym wyżej przykładzie firmy Jaguar Land Rover, gdzie była mowa o zakresie od 140 000 do miliona elementów.

Innym przykładem jest trenowanie systemu rozpoznawania zmęczenia kierowcy, gdzie przygotowany do treningu zbiór danych był stworzony o nagrania ponad 300 osób o zróżnicowanym wyglądzie oddającym zróżnicowanie antropologiczne wyglądu człowieka. Każda z osób była nagrywana przez kilkanaście godzin podczas prowadzenia samochodu w różnych warunkach oświetlenia i pogody. W efekcie powstał zbiór nagrań o rozmiarze ponad 100TB, a dane, ze względu na rozmiar, były transportowane na fizycznych nośnikach.

Dla porównania największym na świecie otwartym zbiorem obrazów jest Google Open Images z 9 milionami zdjęć, podzielonych na 600 klas. Na tych zdjęciach znajduje się 14 milionów obiektów oznaczonych i opisanych przez ludzi, dla wykorzystania w treningu modeli AI.

Wywiady z polskimi firmami wskazują, że do tworzenia swoich rozwiązań, wykorzystują one dane pochodzące od klientów, lub z otwartych zbiorów.

W przypadkach, kiedy są dostępne bardzo duże zbiory danych do ich kategoryzacji ze względu na występujący wzór używa się najczęściej metod nienadzorowanego uczenia (ang. *unsupervised learning*). Może to jednak otwierać nową grupę problemów związanych z określaniem reguł decyzyjnych modeli.

Systemy robotyczne, w tym pojazdy autonomiczne, oraz systemy zrobotyzowanej automatyzacji procesów (RPA – ang. Robotic Process Automation)

Pierwsza generacja systemów RPA nie wykorzystywała rozwiązań AI. Dopiero nadchodząca nowa generacja rozwiązań RPA będzie wykorzystywała sztuczną inteligencję w obszarach: *visual understanding, process understanding, document understanding oraz conversational understanding*. Dotychczas głównym źródłem danych był screen scraping i dane z aplikacji firm.

Inna sytuacja występuje w przypadku **autonomicznych robotów i pojazdów**. Jest to krańcowy przypadek złożoności rozwiązań AI. Autonomiczne pojazdy i roboty wymagają analizy danych w czasie rzeczywistym i wykorzystują bardzo różnorodne źródła danych: kamery, lidary, radary, akcelerometry, moduły GPS, dane z infrastruktury wspierającej. Rozmiar i złożoność zbiorów danych do treningu pojazdów i robotów autonomicznych jest bardzo duża. W Polsce nie prowadzi się całościowych prac związanych z pojazdami autonomicznymi. Otwarte zbiory danych np. największy otwarty zestaw danych dla pojazdów

autonomicznych Berkeley DeepDrive BDD100k zawiera ponad 100,000 filmów wideo z ponad 1,100 godzin jazdy w różnych okresach dnia i warunkach pogodowych.

Wewnętrzne zbiory danych takich firm jak Tesla, Waymo, Uber są wielokrotnie większe i stanowią często jeden z istotnych zasobów firmy.

Z przeglądu powyższych zastosowań sztucznej inteligencji można wysnuć kilka wniosków:

- Polskie startupy AI i firmy AI pracują zwykle na danych dostarczanych przez klientów i na rzecz tych klientów. 78% wskazań w raporcie Digital Poland;
- Polskie firmy pracują przygotowując konkretne rozwiązania dla poszczególnych firm, co nie ułatwia tworzenia kompletnych produktów lub technologii, które mogą być oferowane na szerokim rynku;
- Polskie firmy pracują często na mniejszych zbiorach danych niż ich międzynarodowi konkurenci;
- W Polsce nie ma dużych otwartych zbiorów danych do wykorzystania w treningu modeli AI;

Horyzontalne zasoby danych

Dane geolokalizacyjne

Dane geolokalizacyjne stanowią jedną z najważniejszych kategorii danych wykorzystywanych w gospodarce. Istotność tych zasobów danych jest powszechnie uznana. Już w marcu 2007 roku została uchwalona DYREKTYWA 2007/2/WE PARLAMENTU EUROPEJSKIEGO I RADY ustanawiająca infrastrukturę informacji przestrzennej we Wspólnocie Europejskiej (INSPIRE).

Otwarte dane geolokalizacyjne są istotne z powodu potencjału dla lokalnej gospodarki, równego dostępu do informacji dla różnych grup firm, możliwości tworzenia nowych produktów i usług.

Dodatkowo z perspektywy państwa ważne mogą być kwestie niezależności względem dostawców komercyjnych, kontroli nad kosztem dostępu do takich danych dla obywateli i firm.

W Polsce implementacja dyrektywy jest prowadzona przez Główny Urząd Geodezji i Kartografii, który prowadzi systemy TERYT oraz GEOPORTAL. Systemy te oferują kilka usług głównie skupiając się na wsparciu procesów administracyjnych. Najpopularniejszą usługa

jest usługa danych katastralnych. Na koniec 2016 roku z systemu korzystało niespełna 400 tysięcy użytkowników.

Geoportal i Teryt agregują informacje z wielu źródeł, w tym z powiatów. W wielu przekrojach danych nie ma informacji ze wszystkich powiatów. Dane/mapy są najczęściej udostępniane z myślą o ich wydruku, niewystarczające jest ucyfrowienie i promocja wykorzystania w postaci usług cyfrowych.

Domyślny jest statyczny dostęp i statyczne wykorzystanie danych. W systemach nie ma w zasadzie wzmianek o wykorzystywaniu tych w czasie rzeczywistym przez pojazdy, drony, użytkowników, których interakcje dodają wartości do map.

Chociaż dane geolokalizacyjne są niezwykle istotną kategorią danych dla budowy rozwiązań sztucznej inteligencji obecnie oferowane usługi dostępne w systemach administracji, nie dają alternatywy dla oferowanych przez firmy komercyjne. Olbrzymia większość firm wykorzystuje dane komercyjne ze względu na łatwość dostępu, rozdzielczość i precyzję lokalizacji, łatwość integracji z innymi produktami cyfrowymi.

Raport Open Data Institute „*The UK geospatial data infrastructure: challenges and opportunities*" identyfikuje klika negatywnych konsekwencji uzależnienia gospodarki od map komercyjnych. Są to:

1. Brak kontroli nad zawartością wyświetlanych map i rekomendacjami dla ich użytkowników;
2. Możliwość pogłębiania rozwarstwienie i wykluczenia cyfrowego obszarów nieatrakcyjnych komercyjnie;
3. Uzależnienie tworzenia map i usług od powiązanego z nim zysku;
4. Zagrożenia prywatności związane ze śledzeniem użytkowników korzystających z urządzeń i aplikacji umożliwiających lokalizacje, telefony, zegarki, monitory aktywności, etc.;
5. Możliwości praktyk monopolistycznych i dyktowania cen dla lokalnych użytkowników.

Akceptacja stopnia zależności od produktów komercyjnych jest kwestią wyboru strategii przez państwo. Wydaje się zasadne, aby polepszać jakość dostępnych w Polsce otwartych systemów i ocenić ich przydatność i gotowość z punktu widzenia tworzenia nowoczesnych usług cyfrowych. Istotna sprawą pozostaje edukacja rynku i promocja istniejących zasobów,

które są dalece niewystarczające. Polskie startupy nie traktują danych z systemu Geoportal, jako użytecznego źródła danych dla budowy ich produktów.

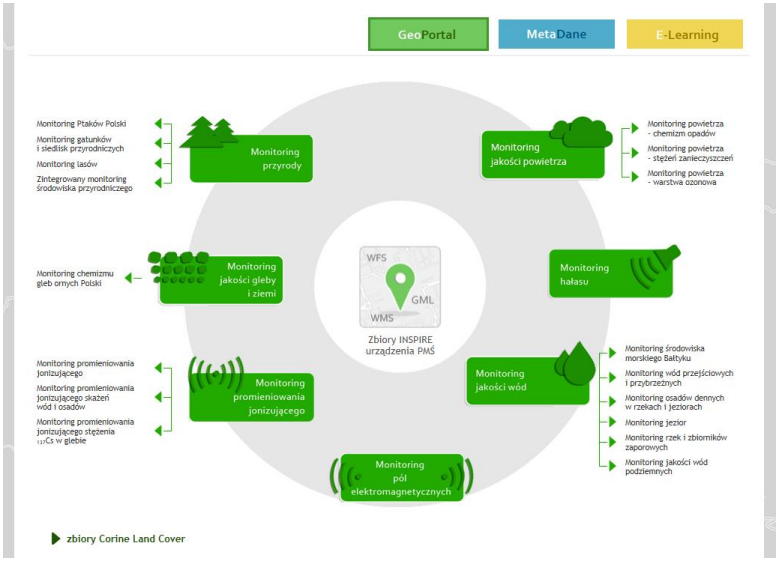
Dane metereologiczne

Instytut Meteorologii i Gospodarki Wodnej oferuje cały szereg danych związanych z prognozami meteorologicznymi poprzez cyfrowe interfejsy API. Dostępne są również dane historyczne dla wybranych miejsc, w których prowadzone są pomiary. IMGW wydaje się być dobrze przygotowane do współpracy przekazywania danych w celu ponownego wykorzystania.

Monitoring środowiska

Za monitorowanie środowiska w kliku przekrojach danych odpowiada Główny Inspektorat Ochrony Środowiska. Oferuje on całe spektrum danych związanych z monitoringiem. Przejrzysty *interface* pozwala na łatwą nawigację w portalu GIOŚ oraz dostęp do danych zgodnie z wytycznymi dyrektywy INSPIRE. Seria danych na portalu inspire.gios.gov.pl, kończą się jednak na roku 2015 i na portalu nie ma referencji do zbiorów danych z kolejnych lat.

GIOŚ uczestniczy również w europejskich projektach *Corine Land Cover* i udostępnia tworzone w ramach tych projektów dane. *Dane Corin Land Cover* dostępne są do 2018 roku.



Rysunek 6. Zbiory Corine Land Cover. Źródło: GIOŚ.

Zasoby języka polskiego

Dostępny jest Narodowy Korpus Języka Polskiego www.nkjp.pl stworzony przez Instytut Podstaw Informatyki PAN, Instytut Języka Polskiego PAN, Wydawnictwa Naukowe PWN oraz Zakład Językoznawstwa Komputerowego i Korpusowego Uniwersytetu Łódzkiego. W ramach tego zasobu dostępny jest referencyjny korpus polszczyzny zawierający ponad 1,5 miliarda słów, wyposażony w dwie dedykowane wyszukiwarki IPI PAN i PELCRA.

Zasoby języka polskiego są również dostępne w Repozytorium Cyfrowym Instytutów Naukowych, gdzie zbiór Instytutu Języka Polskiego PAN zawiera cyfrowe wersje prawie 7300 publikacji.

Narodowy Korpus Języka Polskiego był zbudowany w latach 2008-2012. Wydaje się, że nie ma barier w dostępie do całościowych zasobów języka w formie tekstowej.

Inaczej ma się sprawa dostępu do baz danych nagrań języka. Istnieją zbiory Biblioteki Narodowej zawierające ponad 200 tysięcy nagrań, czy też zbiory Instytutu Sztuki Polskiej Akademii Nauk, ale nie są one przygotowane do wykorzystania w obszarze rozpoznawania mowy. Zbiory nagrań powstałe w działalności gospodarczej są zaś własnością firm i instytucji i z racji ich charakteru nie są udostępniane.

Należy poddać analizie zasadność budowy lub otwarcia na publiczne wykorzystanie modułów rozpoznawania mowy (speach to text) dla języka polskiego. Alternatywą jest wykorzystanie narzędzi komercyjnych oferowanych przez wiodące firmy technologiczne z możliwymi zastrzeżeniami podobnymi do opisanych w rozdziale poświęconym danym geolokalizacyjnym.

Określenie rozmiarów zasobów danych

W oparciu o dostępne informacje nie jest możliwe oszacowanie wielkości zasobów danych w Polskiej gospodarce.

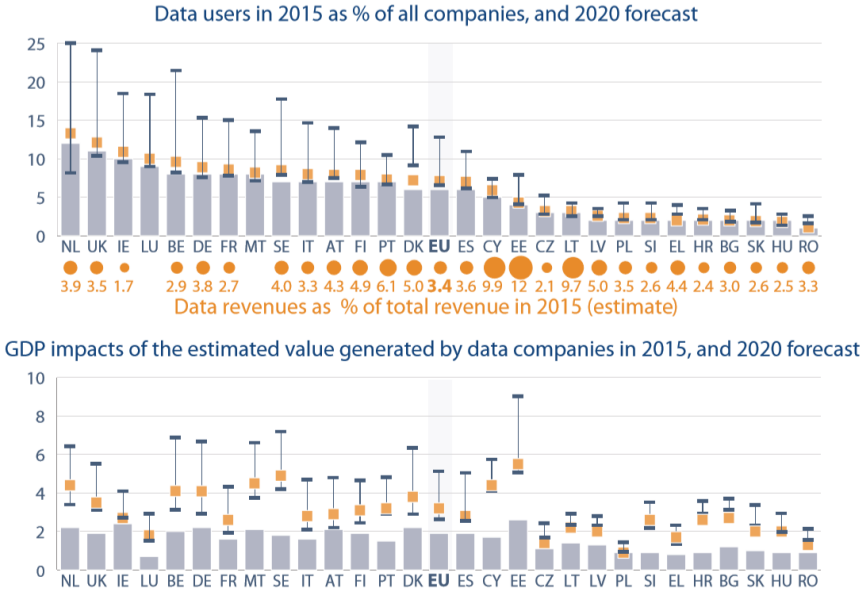
Celowym byłoby zlecenie badania, które oceniłoby zasoby danych w oparciu o wybraną metodykę np. w poszczególnych branżach, samorządach (smart cities) oraz rozmiar danych tworzony w modelach B2C przez polskich użytkowników Internetu i sieci społecznościowych.

Na mniejszy niż w innych gospodarkach dostępny zasób danych w Polsce wskazuje cały szereg informacji.

Polskie startupy jako drugą największa barierę rozwoju wskazują trudności z dostępem do danych. Na ten problem w badaniu przeprowadzonym przez Fundację Digital Poland

wskazuje 38% badanych firm i jest to druga bariera po 41% wskazań dla braku wiedzy korzyściach ze stosowania rozwiązań AI. Wniosek ten jednak, nie znajduje poparcia w wywiadach z firmami i w komentarzach podczas warsztatu DSF.

Badania dla Parlamentu Europejskiego '*Economic impact of Big Data*' (wykonane we wrześniu 2016 roku) wskazują na najniższy udział w PKB firm oferujących usługi analityczne w Polsce w porównaniu do innych krajów Unii Europejskiej, prezentują również pesymistyczną prognozę w tym obszarze.



Rysunek 7. Udział w PKB firm oferujących usługi analityczne Źródło: Badania dla Parlamentu Europejskiego 'Economic impact of Big Data'; wykonane we wrześniu 2016r.

Występują również różnice w ocenie istotności rozwoju infrastruktury sprzętowej dla rozwoju rozwiązań AI. W wielu krajach np. w Anglii dostęp do wydajnej infrastruktury sprzętowej jest postrzegany jako jeden z krytycznych elementów powodzenia w tworzeniu rozwiązań AI (trzeci najważniejszy). Polskie startupy AI nie wskazują dostępu do mocy obliczeniowej jako bariery rozwoju i wskazują, że rozwijają swoje rozwiązania na infrastrukturze o niewielkich mocach obliczeniowych lub w środowiskach chmurowych, co również wskazuje na niezbyt wielki rozmiar wykorzystywanych zbiorów danych.

Przeprowadzone z polskimi firmami AI wywiady potwierdzają, że firmy nastawione są na wykorzystywanie danych klientów, dla których budują rozwiązania AI, jednak w przeciwieństwie do ankiety z badania „Map of the Polish AI" Fundacji Digital Poland, nie traktują dostępu do danych jako bariery. Badanie fundacji Digital Poland, można zatem raczej

odbierać jako przejaw aspiracji polskich firm, które chciałyby tworzyć bardziej skomplikowane rozwiązania, jednak nie mogą ich tworzyć z racji braku danych.
Wydaje się, że obecne zasoby danych są wystarczające dla rozwoju rozwiązań w konkretnych obszarach i dla ograniczonej grupy klientów, którzy takie dane posiadają. Jednocześnie z racji ograniczonej dostępności danych w skali gospodarki trudne będzie zwiększenie wykorzystania rozwiązań AI w stopniu zbliżonym do innych krajów europejskich. Może to również w dłuższej perspektywie ograniczać rozwój polskiej branży AI, która będzie pozbawiona wsparcia z lokalnego rynku zastosowań.

Modele zarządzania danymi – Data Governance

Termin *data governance* używany jest od lat, jednak wraz ze wzrostem wagi danych w procesach zarządzania i podejmowania decyzji, rola data governance stale rośnie.
Data governance określa stworzenie i stosowanie standardów gromadzenia, przetwarzania, przesyłania i przechowywania danych z perspektywy zdefiniowania poszczególnych ról i odpowiedzialności w organizacji, zasad dostępu, kontroli, ochrony danych i bezpieczeństwa. Początkowo odnosił się do zarządzania danymi i polityką danych w przedsiębiorstwie, jednak obecnie definicja data governance ma charakter szerszy i odnosi się do wszystkich relacji przetwarzania danych w zdefiniowanym zakresie: firmy, platformy, branży.
W ramach data governance wydzielić należy klika kluczowych ról:

- Użytkownik danych;
- Właściciel danych;
- Data steward.

Na szczególną uwagę z perspektywy data governance zasługuje Data steward. SAS Institute definiuje jego zadania następująco[1]:

- zarządzanie definicjami danych rozpoczynając od ich znaczenia biznesowego, aż po techniczne określenie struktur i atrybutów opisujących obiekty biznesowe i dopuszczalne zakresy przyjmowanych przez nie wartości;
- tworzenie i realizowanie polityk oraz procedur związanych z zarządzaniem i uzgadnianiem zmian w definicjach danych oraz zmian w systemach informatycznych przetwarzających dane;

[1] https://blogs.sas.com/content/brightdata/2016/05/10

- tworzenie zasad zarządzania jakością danych, których są właścicielami, monitorowanie jakości tych danych oraz realizowanie zadań związanych z podnoszeniem ich jakości (w tym uzgadnianiem konfliktów i czasem wręcz ręczną ich korektą);
- wspieranie właściwego i skutecznego wykorzystania danych przez organizację oraz współpracy w kreowaniu strategii rozwijania środowisk przetwarzania i analizy danych.

W przypadku tworzenia rozwiązań AI **wyzwania data governance rosną z powodu mnogości źródeł i właścicieli danych, różnorodności danych, wielu zaangażowanych organizacji.** Zarządzanie danymi w wykorzystaniu pojazdów autonomicznych jest najlepszym przykładem pokazującym kompleksowość tego zagadnienia. Wynika to zarówno z dużej ilości przetwarzanych danych oraz różnorodności partnerów i uczestników ekosystemu pojazdów autonomicznych (z punktu widzenia rożnych realizacji funkcji w rozwiązaniach oraz z punktu widzenia różnych form organizacyjnych). Zarządzanie danymi w obszarze wykorzystania pojazdów autonomicznych opisuje dokument „*Autonomous driving cityscape - Data governance as an enabler for autonomous driving*" przygotowany przez firmy T-Systems i Detecon Consulting w 2019 roku.

Raport wskazuje na różnorodne źródła: dane z sensorów, dane z samochodów, dane informatycznych systemów miejskich, dane operatorów telekomunikacyjnych, dane kierowców (w tym ich profile) oraz różnorodne kombinacje tych danych.

Wskazuje na wielu interesariuszy: od firm powiązanych z budową i utrzymaniem rozwiązań technicznych, przez władze miejskie i federalne, producentów samochodów, operatorów telekomunikacyjnych, organizacje branżowe, organizacje pozarządowe i ruchy obywatelskie.

Konkluzją raportu jest rekomendacja tworzenia i przeprowadzania projektów pilotażowych, które umożliwią określenia zasad gromadzenia danych i wykorzystywania ich przez zainteresowane tworzeniem nowych aplikacji i rozwiązań firmy.

Te zalecenia wynikają również z przygotowanych przez Komisję Europejską strategicznych dokumentów opublikowanych w maju 2018 roku w ramach pakietu MOBILITY PACKAGE III, które wskazują na potrzeby stworzenia systemu współpracy i wymiany danych z udziałem wszystkich użytkowników dróg i infrastruktury transportowej. Dokument odnosi się do danych generowanych przez wszystkie samochody, wyposażone w urządzenia do

gromadzenia i przesyłania danych, czyli do wielu obecnie jeżdżących samochodów i wszystkich nowych pojazdów. Obecny stan zarządzania tymi danymi przez producentów samochodów, KE uznaje za niewystarczający z punktu widzenia otwarcia danych i budowy nowych aplikacji i rozwiązań. Unijny dokument proponuje również stworzenie cyfrowych platform gromadzenia i wymiany informacji z systemów logistycznych.

European Automobile Manufacturers' Association (ACEA) to stowarzyszenie zrzeszające wszystkich istotnych producentów samochodów[2]. Stowarzyszenie w zgodzie z zaleceniami Mobility Package III rekomenduje stworzenie *extended vehicle data platform servers*, nazywanych również *neutral servers,* platform prowadzonych przez firmy niebędące producentami samochodów, gdzie dane z samochodów, byłyby odpowiednio zarządzane, bezpieczne i dostępne dla firm tworzących cyfrowe usługi i produkty[3].

Programem współdzielenia danych z systemów logistycznych jest AEOLIX (*Architecture for EurOpean Logistics Information eXchange*) http://aeolix.eu , w którym uczestniczą podmioty z kilkunastu krajów europejskich, określając i testując rozwiązania IT dla systemów logistycznych oraz modele zarządzania danymi w takich rozwiązaniach. Niestety podmioty z Polski nie uczestniczą w projekcie AEOLIX.

Wydaje się właściwe i konieczne, aby temat wykorzystania danych z pojazdów, systemów transportowych i logistycznych, został pogłębiony i ujęty w strategii dotyczącej rozwoju rozwiązań sztucznej inteligencji. Stworzone platformy, wirtualne składnice danych, mogłyby służyć za pole doświadczalne dla podobnych inicjatyw w innych branżach.

Brytyjski dokument strategiczny „*GROWING THE ARTIFICIAL INTELLIGENCE INDUSTRY IN THE UK*" przygotowany w 2017 przez prof. Dame Wendy Hall i Jerome'a Pesenti kwestie zwiększania dostępu do danych i tworzenia *data trusts* określa jako pierwszy i jeden z głównych priorytetów strategii rozwoju AI w UK.

Raport zaleca otwieranie zbiorów danych, gdzie to jest możliwe, przygotowywanie ich do maszynowego odczytu oraz przygotowanie zaufanych form dostępu do danych, które z licznych powodów nie mogą być otwarte.

Raport zaleca zwiększenie zakresu współdzielenia danych pomiędzy organizacjami, które dane posiadają, a tymi, które z danych chciałby korzystać, poprzez stworzenie *Data Trusts*,

[2] BMW Group, DAF Trucks, Daimler, Fiat Chrysler Automobiles, Ford of Europe, Hyundai Motor Europe, Iveco, Jaguar Land Rover, PSA Group, Renault Group, Toyota Motor Europe, Volkswagen Group, Volvo Cars, oraz Volvo Group

[3] Źródło: Access to vehicle data for third-party services; ACEA, Grudzień 2016

które są określone jako: umowy i zdefiniowane struktury, aby bezpiecznie i w zaufany sposób umożliwić dostęp do danych i osiąganie założonych korzyści.

Open Data Institute [4] poszerza definicję *data trust* pokazując, kilka różnych znaczeń i aspektów tego terminu, określając data trust jako:

- Organizację zarządzającą i chroniącą dane, które uczestnicy trustu powierzają, kładąc nacisk na realizację interesów właścicieli danych;
- Prawną strukturę zarządzania danymi;
- Zasób/składnicę danych;
- Organizację dostępu do danych z dodaną rolą państwa celem monitorowania i kontroli procesu współdzielenia danych.

Przeniesienie koncepcji data trust/wirtualnej składnicy danych na grunt polski wymaga przezwyciężenia wielu barier.

Dane powinny być zbierane w określonym celu i z określonym kontekstem metadanych, podczas gdy w polskich firmach brakuje często świadomości i kompetencji do określenia polityki danych oraz nie ma zdefiniowanych podstaw współpracy z firmami, które dane chciałby wykorzystać. Nie ma również przykładów umów i porozumień regulujących zasady funkcjonowania wirtualnych składnic danych.

Przezwyciężenie tych ograniczeń może być łatwiejsze w wybranych branżach, gdzie współdzielenie danych jest mniej kontrowersyjne i gdzie istnieje większa świadomość potencjalnych korzyści. Jest to również obszar, w którym państwo powinno wspierać działania edukacyjne i wspierać programy pilotażowe.

Budowa świadomości korzyści, promowanie przykładów dobrze działających wdrożeń oraz brak informacji na temat źródeł danych są powtarzającymi się problemami ekosystemu budowy i wykorzystania rozwiązań AI. Podkreśla to znaczenie działań edukacyjnych i promocyjnych w tym zakresie.

W przypadku kiedy firma określa dane jako szczególnie istotne z perspektywy utrzymania konkurencyjności, może zbudować zamknięte środowisko, w którym udostępnia dane dla wybranej grupy firm, specjalizujących się w tworzeniu rozwiązań AI.

Wg podobnego schematu działa inicjatywa Data Pitch (https://datapitch.eu) prowadzona w ramach programu Horizon 2020. W przypadku tego programu zarządzaniem danymi, jako

[4] theodi.org/article/what-is-a-data-trust/

data steward zajmuje się *University of Southampton*, a wybrane startupy mogą się skupić nad rozwiązywaniem problemów biznesowych zdefiniowanych przez Data Pitch.

„I" – Issues – Bariery

Segment firm rozwijających rozwiązania AI w Polsce rośnie szybko bazując na sprzedaży kompetencji i znajdując klientów w większości poza granicami Polski. Firmy te rozwijają swoje rozwiązania w oparciu o dane klientów. W Polsce również występują grupy klientów, firm stosujących rozwiązania AI, które posiadają zasoby danych potrzebne do rozwoju rozwiązań AI.

Sytuacja taka występuje w kliku sektorach (np.: finansowym, retail, call centers, telekomunikacyjnym, e-commerce), które z powodów regulacyjnych lub z powodu natury modeli biznesowych gromadziły dane, które mogą być wewnętrznie wykorzystywane.

Nie ma jednak takich firm wiele, oraz z reguły posiadają one mniejsze zasoby danych niż ich odpowiedniki z dużych rynków, co w niektórych przypadkach może ograniczać konkurencyjność stosowanych przez nie rozwiązań. Nie jest również jasne czy firmy te sięgają do większych zasobów danych pochodzących z poza tych przedsiębiorstw.

Pozwala to na rozwój ekosystemu rozwiązań AI na wstępnym etapie, nie daje jednak gwarancji osiągnięcia sukcesu w dłuższej perspektywie, kiedy zacznie następować konsolidacja rozwiązań AI i firm je oferujących.

Zidentyfikowaliśmy następujące bariery ograniczające tempo i skale wdrożeń AI w gospodarce, które również przekładają się na ograniczenia rozwoju firm AI.

1. Brak zrozumienia roli danych w gospodarce
    a. Niewystarczające badania dostępności danych w polskiej gospodarce, zwłaszcza zasobów danych przydatnych do maszynowego odczytu i wykorzystania w rozwiązaniach AI;
    b. Brak badań rozmiarów zasobów danych w polskiej gospodarce, w tym w firmach państwowych i prywatnych;
    c. Brak chęci budowy przewagi konkurencyjnej w oparciu o dane;
    d. Brak promocji i niespójne informacje o dostępie do danych w systemach publicznych;

   e. Brak programów promujących współpracę w obszarach współdzielenia danych i uczących różnych uczestników przełamywać bariery komunikacji i braku współpracy pomiędzy branżami;
   f. Niewystarczająca liczba programów pilotażowych, eksperymentujących z budową nowych modeli biznesowych opartych o dane.
2. Niezrozumienie specyfiki rodzajów i formatów danych potrzebnych dla budowy i wykorzystania rozwiązań AI.
3. Mniejsza skala danych gromadzonych w gospodarce, przedsiębiorstwach oraz otwartych zbiorach danych. Wykorzystywanie przez polskie firmy AI mniejszych zbiorów danych w porównaniu do ich międzynarodowej konkurencji.
4. Ograniczanie zasad *data governance* do danych wewnętrznych i danych otwartych.
5. Brak zasad *data governance*: polityki gromadzenia, przetwarzania i udostępniania dla danych współdzielonych:
   a. Brak definicji wirtualnych składnic danych i zasad ich funkcjonowania;
   b. Brak wybranych branż, określenia celów, korzyści i inicjacji programów pilotażowych oraz odpowiedzialności za te działania;
   c. Brak przykładów i wzorców organizacji współdzielenia danych w wirtualnych składnicach danych.
6. Generyczne podejście do rozwiązań sztucznej inteligencji bez zrozumienia specyfiki ich zastosowań:
   a. Bariery rozwoju systemów analitycznych, predykcyjnych i optymalizacyjnych:
      i. Dobra sytuacja w branży finansowej pomimo prawnych ograniczeń dostępu do danych i ich współdzielenia w oparciu o zamknięte dane przedsiębiorstw;
      ii. Dobra sytuacja w handlu tradycyjnym i internetowym w oparciu o zamknięte dane przedsiębiorstw;
      iii. Brak polityki wykorzystywania danych w samorządach w obszarach smart city;
      iv. Niewielka liczba programów pilotażowych i niewystarczająca współpraca z wybranymi kluczowymi branżami w celu realizacji programów pilotażowych (telekomunikacja, motoryzacja);
      v. Brak jasnej strategii w pozostałych branżach np.: medycznej, rolnictwie.

b. Bariery rozwoju systemów NLP i chatbotów:
    i. Brak otwartych narzędzi rozpoznawania mowy (speach to text) dla języka polskiego;
    ii. Osłabiona pozycja konkurencyjna polskich firm w odniesieniu do międzynarodowych dostawców, oferujących moduły speach to text;
    iii. Brak zdefiniowanej polityki w zakresie wykorzystywania rozwiązań komercyjnych.
c. Bariery rozwoju systemów rozpoznawania obrazów oraz systemów analizy wideo i widzenia maszynowego:
    i. Mniejsza skala zbiorów danych wykorzystywanych przez polskie startupy;
    ii. Otwarte międzynarodowe zbiory obrazów, mogą narażać modele na bias (skrzywienie rezultatów) z powodu niedostatecznego oddania charakterystyki populacji danych obiektów w Polsce.
d. Bariery rozwoju systemów robotycznych, w tym pojazdy autonomicznych, oraz systemów RPA:
    i. Nie zidentyfikowano barier dla rozwoju systemów RPA;
    ii. Brak współpracy z firmami motoryzacyjnymi w obszarze testowanie pojazdów autonomicznych, budowy infrastruktury ich wsparcia i wykorzystywania danych pochodzących z tych obszarów, pomimo silnej obecności branży motoryzacyjnej w Polsce;
    iii. Brak informacji na temat wykorzystywania danych systemów przemysłowych i ich współdzielenia.

7. Niespójna polityka w obszarze danych horyzontalnych istotnych dla państwa, społeczeństwa i gospodarki:
    a. Niejasna polityka definiująca znaczenie i zasady utrzymywania otwartych horyzontalnych zbiorów danych:
        i. Efektywniejsza tam gdzie występują regulacje europejskie np. INSPIRE, słabsza w innych obszarach;
        ii. Niejasne zasady koordynacji prac i odpowiedzialności w obszarze przygotowania danych cyfrowych;

iii. Rosnąca zależność od rozwiązań komercyjnych w obszarze wykorzystania dokładnych map, brak wiedzy o zasadach wykorzystywania zasobów publicznych;
iv. Brak otwartych narzędzi rozpoznawania mowy (speach to text) dla języka polskiego;
v. Brak zdefiniowanej polityki w zakresie wykorzystywania rozwiązań komercyjnych.

b. Brak lub ograniczone przygotowanie otwarcia zbiorów danych dla odczytu maszynowego, brak interface'ów API w niektórych serwisach państwowych.
c. Brak aktualizacji zbiorów danych, wiele projektów zostało zarzuconych po zakończeniu finansowania z środków unijnych.

8. Niewystarczający transfer wiedzy i dobrych praktyk do Polski z:
   a. firm międzynarodowych działających w Polsce;
   b. niektórych projektów Unii Europejskiej (Data Pitch, AEOLIX).

"R" – Recommendations – Zalecenia

Rozwiązań, dla przyspieszenia rozwoju gospodarki i systemów AI w oparciu o dane, należy szukać w kilku obszarach. Są one równie ważne:

- Diagnoza dostępności i wykorzystania danych w Polskiej gospodarce;
- Definicja zasad i budowa praktyk data governance w Polsce, zwłaszcza w odniesieniu do danych współdzielonych;
- Rozwój dostępnych dla AI zasobów danych;
- Edukacja rynku w zakresie dostępności danych, specyfiki rozwiązań AI zasad data governance i modeli biznesowych opartych o szersze wykorzystanie danych;

Z tej perspektywy oraz wychodząc naprzeciw wyzwaniom sformułowanym w sekcji „Bariery" uważamy za priorytetowe, aby zrealizować następujące zalecenia:

1. Przeprowadzić badanie zasobów danych do wykorzystania w polskiej gospodarce dla rozwiązań AI:

   a. Potrzeby związane z wykorzystaniem danych w systemach AI w poszczególnych obszarach stosowania AI – badanie potrzeb firm AI i klientów końcowych (B2B);
   b. Potrzeby związane z gromadzeniem, udostępnianiem i wykorzystywaniem danych w poszczególnych segmentach gospodarki i administracji, z uwzględnieniem samorządów (np. przemysł, rolnictwo, służba zdrowia, transport, transport publiczny, bezpieczeństwo publiczne, etc.);
   c. Gotowości firm na szersze wykorzystywanie narzędzi analitycznych i rozwiązań AI oraz barier w tym obszarze.
2. Sformułować diagnozę wykorzystania danych w gospodarce, wykorzystując zasoby Fundacji Platformy Przemysłu Przyszłości, poddać ją publicznej dyskusji i przeprowadzić rewizję strategii w oparciu o wnioski.
3. Doprecyzować założenia polityki względem wykorzystania danych horyzontalnych dla budowy rozwiązań AI w Polsce oraz określenia stopnia wymaganej niezależności od rozwiązań komercyjnych:
   a. Podjąć decyzję w zakresie budowy i utrzymania narzędzi rozpoznawania mowy (speach to text) dla języka polskiego;
   b. Zdefiniować zasady aktualizacji i utrzymania zasobów w tych obszarach z punktu widzenia wymagań odczytu maszynowego i budowy rozwiązań AI;
   c. Sformułować maksymalnie jednolite i spójne zasady dostępu do tych danych, wraz z zasadami promocji i informacji o tych zasadach.
4. Przypisać do *Digital Innovation Hubów* (DIH) zadania związane z aktywacją rozwiązań pilotażowych z:
   a. przypisaniem odpowiedzialności za kluczowe segmenty (rolnictwo, przemysł, telekomunikacja, transport – w tym transport autonomiczny, służba zdrowia, etc);
   b. włączeniem wszystkich kluczowych partnerów (biznesu, nauki, administracji i samorządów) i dbałością w współpracę;
   c. dbałością o współpracę pomiędzy partnerami i otwarte publikowanie wyników i wniosków, gdzie to możliwe.
5. Przypisać do DIH odpowiedzialność za przygotowanie rekomendacji data governance dla danych otwartych i współdzielonych oraz za stworzenie testowych wirtualnych

składnic danych (z wykorzystaniem infrastruktury publicznej – centrów obliczeniowych HPC, , lub prywatnej)

6. Wprowadzić temat prowadzenia programów pilotażowych i dzielenia się (gdzie to możliwe) danymi i wnioskami powstałymi w działaniu tych programów w dyskusjach z przedsiębiorstwami zagranicznymi w Polsce celem transferu dobrych praktyk i wiedzy z bardziej zaawansowanych gospodarek:
    a. Zainicjować dyskusje z firmami motoryzacyjnymi obecnymi w Polsce na temat dostępu do danych z wyprodukowanych przez nich pojazdów poruszających się w Polsce w odniesieniu do Mobility Package III i zaleceń ACEA;
    b. Poszerzyć zakres pilotów 5G na obszar zastosowań rozwiązań wykorzystujących 5G w różnych gałęziach gospodarki z naciskiem na współdzielenie danych i określanie zasad data governance.
7. Zrobić przegląd programów Unii Europejskiej nastawionych na projekty współdzielenia danych i ocenić zaangażowanie polskich firm, uczelni i instytucji badawczych w te projekty:
    a. Zainicjować dyskusję i zaangażowanie odpowiednich ośrodków badawczych i firm w strategicznie ważne projekty np. AEOLIX;
    b. Ocenić potencjał programu Data Pitch i wykorzystać wnioski dla stworzenia polskiej wersji programu.