ChatGPT

Metodyka klasyfikacji i oceny opłacalności zbiorów danych dla AI

Definicje kluczowych pojęć

Opłacalność ekonomiczna – W kontekście zbiorów danych do trenowania AI oznacza zdolność projektu związanego z danymi do wygenerowania wartości (przychodów, oszczędności) przewyższającej koszty w określonym horyzoncie. Formułuje się to zwykle poprzez kryterium **NPV ≥ 0** (net present value), czyli sumę zdyskontowanych przepływów pieniężnych minus koszt początkowy (CAPEX) [1]. Innymi słowy, inwestycja w dane jest uznana za opłacalną, jeśli bieżąca wartość przyszłych zysków pokrywa nakłady inwestycyjne i operacyjne. Standardowo definiuje się także **ROI** (stopa zwrotu z inwestycji) jako stosunek zysku netto do poniesionych kosztów [2]. Dla projektów danych wysokiego ryzyka istotna jest również wartość oczekiwana ryzyka błędu – **EV(błędu)** – którą szacuje się jako prawdopodobieństwo błędu pomnożone przez koszt takiego błędu [2]. Redukcja tego oczekiwanego kosztu (ΔEV) dzięki lepszym danym lub nadzorowi może stanowić wymierną **wartość dodaną** systemu.

CAPEX vs OPEX – **CAPEX** (nakłady inwestycyjne) to jednorazowe koszty utworzenia zbioru danych lub infrastruktury (np. zakup sensorów, sprzętu, budowa habitatów orbitalnych). **OPEX** (koszty operacyjne) to bieżące wydatki na utrzymanie pozyskiwania i przetwarzania danych (np. koszty załogi, zasilania, łączności, uaktualniania danych). Przykładowo, **orbitalna stacja badawcza** ma CAPEX rzędu ~$3 mld USD (wg prospektu Starlab) [3], zaś jej roczne OPEX mogą sięgać miliardów – utrzymanie Międzynarodowej Stacji Kosmicznej (ISS) kosztuje ok. **4,1 mld USD rocznie** [4]. To pokazuje, że aby przedsięwzięcie o dużym CAPEX i OPEX było opłacalne, musi generować ogromne strumienie wartości lub przychodów. W warunkach orbitalnych istnieją tzw. **ceny cienia** kluczowych zasobów: wyniesienie ładunku na orbitę ~20 000 USD/kg, sprowadzenie na Ziemię 40 000 USD/kg, utylizacja śmieci 20 000 USD/kg, a roboczogodzina astronauty ok. **130 000 USD** [5]. Te wartości odzwierciedlają ukryty koszt pozyskania danych w tak ekstremalnym środowisku i muszą być wkalkulowane w model biznesowy zbioru danych.

Produkt danych vs dane surowe – **Dane surowe** to nieprzetworzone informacje (np. surowe logi sensorów lub tekst bez adnotacji). **Produkt danych** to opracowany zbiór danych wzbogacony i gotowy do użycia: oczyszczony, opisany, często połączony z narzędziami i metadanymi zwiększającymi jego wartość. W nowoczesnym podejściu podkreśla się, że dane powinny być traktowane jak produkt, z odpowiednią dokumentacją i gwarancją jakości. Standard **FAIR** (Findable, Accessible, Interoperable, Reusable) definiuje reżim zarządzania danymi tak, by były łatwo znajdowalne, dostępne, interoperacyjne i podatne na ponowne wykorzystanie – co jest spójne z ideą dostarczania **produktów danych zamiast surowych logów** [6]. Ponadto zaleca się stosowanie standardów **Datasheets for Datasets** (arkusze danych opisujące motywacje, skład, proces powstawania i użycie zbioru danych) oraz **Model Cards** (karty informacyjne modeli AI, ujawniające zamierzone zastosowania, wyniki i ograniczenia modelu) celem zapewnienia transparentności i zaufania do oferowanych danych i modeli [7] [8]. Taki **pakiet danych** (dataset + dokumentacja + rodowód + polityka dostępu, a dla modeli wytrenowanych na tych danych także model card) zwiększa wartość rynkową – bez audytu i kontekstu nawet unikalny zbiór traci na wartości [9] [8].

Dane brakujące (rzadkie) vs dane nasycone (commodity) – Kluczowa koncepcja **rynku danych** mówi, że nie każda informacja ma wysoką cenę. **Opłacalność pochodzi z danych, których brakuje** – czyli

rzadkich, trudno zastępowalnych i rozwiązujących pilny problem o dużej wartości – **a nie z posiadania dowolnych dużych wolumenów danych** [10]. Dane powszechnie dostępne stają się towarem o cenie bliskiej kosztu krańcowego. Wysoką wartość utrzymują tylko te zbiory, które oferują unikalny wgląd zanim rynek się nasyci lub pojawią się substytuty [11]. **Marginalny zwrot z danych** maleje wraz z ich upowszechnianiem – pierwsze takie dane mogą dać przewagę konkurencyjną lub znacząco poprawić model AI, ale każda kolejna partia danych tego samego typu daje coraz mniejszą poprawę jakości (prawo malejących przychodów z kolejnych danych). Dlatego w ocenie opłacalności zbioru należy uwzględnić **ryzyko nasycenia rynku**: czy w przyszłości podobne dane staną się łatwo dostępne, obniżając cenę? Przykładowo, w raporcie wskazano, że tylko dane **rzadkie i trudno zastępowalne utrzymują wysoką cenę w czasie**, zanim konkurencja i dostępność nie zbiją ceny do poziomu kosztu krańcowego [12].

Koszt błędu – W wielu zastosowaniach AI kluczowym komponentem wartości jest unikanie błędów. **Koszt błędu** to finansowe (lub społeczne) skutki popełnienia błędnej decyzji na podstawie danych lub modelu. Może to być np. szkoda w sprzęcie wskutek złej predykcji, błąd medyczny, niewykrycie oszustwa, kolizja autonomicznego pojazdu – każde z tych zdarzeń ma wymierny koszt (strata finansowa, kara, odpowiedzialność prawna, utrata reputacji). W systemach **Human-AI-In-the-Loop (HITL)** podkreśla się, że opłacalność takiego układu szczególnie zależy od **istotności kosztu potencjalnego błędu** – jeśli koszt pomyłki jest wysoki, inwestycja w dodatkowe dane, lepszy model lub nadzór człowieka może być w pełni uzasadniona ekonomicznie [13] [14]. Wartość danych przejawia się tu jako **redukcja oczekiwanej straty**: lepszy zbiór danych może obniżyć prawdopodobieństwo błędu lub jego skutki (np. model medyczny uczulony na rzadkie przypadki zmniejsza ryzyko kosztownej pomyłki diagnostycznej). W rachunku ekonomicznym ujmuje się to poprzez wspomnianą metrykę EV(ryzyka) = P(błąd) * koszt(błędu) i obserwuje, o ile EV spada po wdrożeniu ulepszonych danych/modelu [15] [16].

Architektura metodyczna oceny zbioru danych

Proponowana metodyka to **wieloetapowy framework** łączący kryteria techniczne, ekonomiczne i organizacyjne. Jego celem jest **usystematyzowanie oceny potencjalnego zbioru danych** – od charakterystyki samego zbioru, przez analizę jego wartości rynkowej, po koszty i sposób wdrożenia w organizacji. Poniżej przedstawiono kluczowe komponenty tej architektury:

1. **Identyfikacja i klasyfikacja zbioru danych** – Na początek określamy, z jakim typem danych mamy do czynienia i w jakim kontekście będą one użyte. Czy są to dane surowe czy już przetworzone produkty? Czy dotyczą krytycznych systemów (**safety-critical**), danych biologicznych/medycznych, danych z systemów autonomicznych, czy może danych czysto cyfrowych? Każdy typ będzie miał inny profil ryzyka i wartości (patrz następna sekcja klasyfikacji) [17]. Ważne jest też określenie, **czy dane stanowią „brakujący element" na rynku** – innowacyjne i rzadkie – czy też konkurują z wieloma już dostępnymi zbiorami (co sugeruje nasycenie rynku) [12].

2. **Analiza problemu i wartości użytkowej** – Następnie oceniamy, **jaki problem te dane pomagają rozwiązać** i kto jest ich potencjalnym odbiorcą. Popyt na dane wynika z chęci rozwiązania konkretnego problemu o wysokiej stawce, nie z samych danych per se [18]. Dlatego trzeba zrozumieć kontekst domenowy: czy dane te umożliwią np. poprawę bezpieczeństwa systemu, zwiększą autonomię procesu, usprawnią decyzje w środowisku zamkniętym? **Wartość marginalna danych** powinna być oceniona: np. o ile zwiększa się skuteczność modelu AI po dodaniu tych danych? Czy istnieją benchmarki lub testy, gdzie można wykazać przewagę dzięki temu zbiorowi? Na tym etapie warto zaplanować **walidację** – np. poprzez utworzenie prototypowego modelu AI z tymi danymi lub przeprowadzenie eksperymentu porównawczego.

W praktyce zaleca się **szybkie wytworzenie Minimalnego Produktu Danych (MVP-Data)** oraz **dowodu użyteczności**, zanim zainwestuje się na większą skalę [19]. Taki MVP-Data powinien być już przetestowany np. w formie eksperymentu lub mini-benchmarku, aby zmierzyć realną poprawę metryk (np. dokładności modelu, spadku błędów) [20].

3. **Ocena wartości rynkowej i potencjału** – Mając wstępne wyniki, oceniamy potencjał rynkowy zbioru. **Wartość rynkowa** nie jest absolutna – zależy od liczby potencjalnych klientów, alternatyw i trendów. Należy oszacować, **kto zapłaci za te dane** i ile. Czy klientami będą podmioty komercyjne (np. firmy technologiczne, sektor prywatny) czy instytucjonalne (np. agencje rządowe, wojsko)? Struktura popytu decyduje o stabilności przychodów [21] [22]. Przykładowo, dane satelitarne sprzedawane jako usługa subskrypcyjna generują dziś przychody rzędu **0,1–0,25 mld USD rocznie** dla wiodących firm (Planet Labs: ~$244 mln; Spire Global: ~$110 mln; BlackSky: ~$102 mln) [23]. To istotny punkt odniesienia – pokazuje skalę możliwych przychodów z danych kosmicznych. Jednak należy zauważyć, że modele czysto **„data-only"** tych firm działają bez załogi (bezzałogowe satelity), co oznacza znacznie niższe koszty operacyjne niż w przypadku infrastruktury z ludźmi [24] [25]. Jeśli nasz rozważany zbiór danych wymaga drogiej infrastruktury (np. habitat z załogą generujący dane), to sam rynek danych na poziomie setek milionów USD może nie wystarczyć do pokrycia kosztów – potrzebne będą dodatkowe strumienie wartości lub dofinansowanie (ten wątek rozwinięto w przykładzie habitatu orbitalnego). W ocenie potencjału warto też sprawdzić, czy dane te mogą stać się **elementem standardowych benchmarków** w branży – to podnosi ich reputację i popyt. Jeśli np. zbiór posłuży do utworzenia nowego testu dla autonomicznych robotów lub modelu biomedycznego, może zyskać status referencyjny i przyciągnąć użytkowników.

4. **Analiza kosztów pozyskania i utrzymania danych** – Kolejnym elementem jest **szczegółowy rachunek kosztów**: jak trudno i drogo jest zdobyć, przetworzyć oraz utrzymać jakość tego zbioru? Należy wyszczególnić:

5. **Koszt pozyskania**: czy wymaga specjalistycznego sprzętu (np. satelitów, sensorów LIDAR), drogiego eksperymentu, czasu eksperckiego? Przykładowo, dane z habitatów orbitalnych pociągają za sobą **logistykę kosmiczną** – każdy kilogram materiału do stacji to ok. 20 000 USD kosztu wyniesienia [5]. Jeśli dane wymagają udziału ludzi (np. astronautów zbierających próbki), trzeba uwzględnić koszt ich czasu pracy (ISS wycenia godzinę astronauty na ~$130k) [26]. Dla danych naziemnych koszt pozyskania może obejmować zakup urządzeń pomiarowych, wynagrodzenia labelerów (jeśli dane wymagają ręcznego oznaczania), koszty podróży na miejsce zbierania danych itp.
6. **Koszt przetworzenia**: obejmuje przygotowanie danych do użytku – czyszczenie, anonimizację, transformacje, integrację z innymi źródłami. W przypadku dużych zbiorów znaczące są koszty infrastruktury IT (przechowywanie, obliczenia). Jeśli dane są strumieniowe, w grę wchodzi koszt pipeline'u przetwarzania w czasie rzeczywistym.

7. **Koszt kontroli jakości**: zapewnienie, że dane są poprawne, kompletne i aktualne. Tu wchodzą procedury walidacji, testy, audyty oraz **nadzór człowieka** nad danymi. Wysoka jakość jest szczególnie ważna dla **danych safety-critical** (błędy mogą powodować wypadki) oraz danych trenowanych na modelach mających wpływ na życie/zdrowie. Trzeba ocenić, czy jakość można zagwarantować automatycznie, czy potrzebna jest kosztowna weryfikacja manualna. Istnieją wzorce projektowe pozwalające ograniczyć koszt nadzoru – np. **system bramek jakości (go/no-go)** i losowej kontroli, by wcześnie wyłapywać błędy [27] [28]. Badania wskazują, że **koszt nadzoru bywa nieliniowy**: źle zaprojektowany interfejs czy procedury mogą prowadzić do zjawisk jak *automation bias* (ślepe ufanie AI) albo nadmierna weryfikacja, które paradoksalnie **zwiększają koszty** i wydłużają proces [29]. Dlatego w architekturze należy przewidzieć **tanie i mierzalne**

mechanizmy kontroli jakości – np. progi decyzyjne, gating, sampling danych do sprawdzenia – tak, by człowiek interweniował tylko tam, gdzie to konieczne [14].

8. **Zastosowanie standardów i governance** – Metodyka zakłada, że **zarządzanie danymi (data governance)** jest integralną częścią wartości i kosztu zbioru. Dane wysokiej wartości często wymagają spełnienia surowych norm etycznych, prawnych i jakościowych. Należy zaplanować zgodność ze standardami **FAIR** i wdrożenie artefaktów typu datasheet i model card (opisanych w sekcji Definicje) dla danego zbioru [7] [8]. To generuje dodatkowy koszt (czas ekspertów na dokumentację, procedury audytu), ale jest też **elementem przewagi konkurencyjnej** – zaufanie do danych staje się „walutą" na rynku AI [9]. Bez odpowiedniej dokumentacji i demonstracji jakości wiele instytucji nie kupi ani nie wykorzysta danych, zwłaszcza w domenach **wysokiego ryzyka** (np. medyczne AI, autonomiczne pojazdy). Governance obejmuje także polityki prywatności i bezpieczeństwa danych osobowych. Jeśli zbiór zawiera dane wrażliwe (np. biometryczne, behawioralne), trzeba uwzględnić koszty ich ochrony i zgodności z regulacjami (np. GDPR). W środowisku ekstremalnym jak habitat kosmiczny dochodzą czynniki **Human Factors** – NASA wskazuje ryzyka behawioralne izolacji i stresu, co oznacza konieczność dbałości o **bezpieczeństwo behawioralne** i prywatność załogi [30]. Incydenty naruszenia danych czy konflikt w załodze to też koszty ryzyka, które governance ma minimalizować.

9. **Model ekonomiczny opłacalności** – Dysponując powyższymi danymi (wartość i potencjał rynkowy vs koszty i ryzyka), można zbudować **model finansowy** przedsięwzięcia danych. Obejmuje on projekcję przychodów (np. ze sprzedaży licencji na dane, z usług analitycznych nad danymi, z kontraktów na dostęp do platformy badawczej) oraz wydatków (CAPEX na start, coroczny OPEX). Należy policzyć oczekiwane przepływy pieniężne w czasie i zastosować np. zdyskontowane przepływy, by ocenić NPV. **Warunek rentowności NPV $\geq$ 0** musi zostać spełniony, co dla danych projektów oznacza wyznaczenie **progu przychodów** potrzebnych rocznie [1]. Jeśli planujemy stały roczny cashflow, to próg rentowności można wyliczyć z warunku NPV w postaci: sumaryczne przychody roczne > (CAPEX * dyskont) + OPEX [31]. W praktyce np. dla CAPEX ~3 mld USD i kosztu kapitału 12% rocznie, sam **dodatek przychodów** potrzebny ponad pokrycie OPEX to ok. 0,5–0,6 mld USD rocznie [32]. Taka kalkulacja uzmysławia skalę wyzwań – jeśli nasz zbiór danych wymaga takiego CAPEX, musimy celować w setki milionów USD przychodu rocznie, by projekt miał sens. Można też obliczyć **ROI** dla porównania z alternatywnymi inwestycjami: ROI = (wartość z danych – koszt danych) / koszt danych [2]. Użyteczny bywa też rachunek **EV(ryzyka)**: ile kosztują potencjalne szkody przed vs po wdrożeniu danych (ΔEV to oszczędność dzięki danym) [16]. W systemach gdzie celem danych jest redukcja ryzyka (np. zapobieganie awariom), ten *risk-adjusted return* może być głównym uzasadnieniem projektu [33]. Warto podkreślić, że **opłacalność ma wymiar dynamiczny** – trzeba analizować scenariusze: np. co jeśli rynek urośnie/zmaleje, co jeśli pojawi się konkurencja (spadek cen), co jeśli koszty technologii spadną? Metodyka powinna uwzględniać **analizę wrażliwości** modelu na takie zmiany.

10. **Skalowalność i granice wzrostu** – Ostatnim elementem oceny jest spojrzenie strategiczne: czy ten zbiór danych i model biznesowy da się skalować i jak długo utrzyma przewagę? Tu ponownie wraca kwestia **nasycenia rynku danych**. Jeśli opiera się on na danych rzadkich dziś, należy ocenić, czy nie staną się one powszechne jutro. Jeżeli spodziewany jest **efekt sytości** (przesycenia) – np. po zebraniu pewnej ilości danych dalsze ich jednostki nie zwiększają już jakości modelu lub na rynku pojawi się wiele podobnych ofert – to może ograniczyć długoterminowe przychody. Należy wtedy wykazać, że model biznesowy ma mechanizmy obronne: np. stale generuje **nowe unikalne dane** (posuwa granicę eksploracji), dywersyfikuje zastosowania lub ma lojalnych klientów na abonament. W raportach wskazano, że **trwały model** musi być odporny na spadek cen danych – np. dzięki temu, że jeden strumień przychodów ma

charakter stały/kontraktowy i pokrywa koszty bazowe, podczas gdy inne strumienie generują wzrost [34]. Taka dywersyfikacja (np. jeden duży klient instytucjonalny jako *anchor* plus wielu mniejszych na dodatkowe usługi) zabezpiecza projekt przed załamaniem, gdy czysty rynek danych stanieje [35] [36].

11. **Osadzenie w architekturze organizacyjnej** – Metodyka nie kończy się na kalkulacjach – musi być wdrożona poprzez odpowiednią architekturę procesów i ról w organizacji. Chodzi o to, by zebranie i wykorzystanie danych było **skutecznie zintegrowane z ludźmi i infrastrukturą**. Wiele projektów AI zawodzi nie z braku danych, lecz z braku mechanizmów wykorzystania ich potencjału na co dzień (lub kontroli jakości). Dlatego zalecane jest zastosowanie **pętli decyzyjnych z człowiekiem w pętli (HITL)** tam, gdzie to potrzebne – człowiek jako nadzorca albo dostarczyciel informacji zwrotnej powinien być wpisany w proces AI, szczególnie w punktach krytycznych decyzji [37] [38]. W praktyce oznacza to np. bramki zatwierdzające decyzje AI w systemach autonomicznych (gdy wykrywane są anomalie, decyduje człowiek), mechanizmy *override/stop* w systemach sterowania (człowiek może przerwać działanie AI, co w architekturze **swarm** przewidziano przy dronach – decyzje stop są naturalnym miejscem interwencji człowieka [38]), czy choćby losowe audyty rekomendacji AI. Ponadto organizacja powinna adoptować **model „mrowiska"** – w kulturze pracy przypominający kolonię mrówek, gdzie role są płynne i wszyscy dokładają się do wspólnego celu. W kontekście projektów AI i danych oznacza to **wielofunkcyjne zespoły**, w których ludzie mogą pełnić różne role (np. inżynier danych, analityk, operator, tester) w zależności od potrzeby, a system szkolenia pozwala im przemieszczać się między zadaniami [39]. Taka elastyczność jest szczególnie ważna np. w małej załogowej bazie kosmicznej, gdzie każdy członek załogi musi znać się na wielu aspektach (przetwarzanie danych, obsługa urządzeń, podstawy analizy), aby zapewnić odporność systemu na wypadek rotacji lub awarii personelu [39]. Z punktu widzenia wartości ekonomicznej, **redukcja silosów i usprawnienie współpracy Human-AI** skraca cykle pracy i obniża koszty błędów (problemy rozwiązywane są szybciej, zanim urosną w kosztowne awarie) [40].

12. **Pipeline artefaktów i obserwowalność** – Na poziomie technicznym architektura metodyczna zakłada budowę **„kręgosłupa" pipeline'u danych i modeli**, który łączy mechanizmy zbierania danych, ich walidacji oraz ciągłego monitoringu jakości. W raportach opisano to jako ciąg od surowych **„śladów" (traces)** do **produktu danych** z wbudowanymi punktami kontrolnymi (stigmergia + gating) [41]. Wzorując się na rozwiązaniach open-source, można przyjąć następujący schemat:

13. **Automatyczne zbieranie śladów** – wszystkie zdarzenia, zmiany i obserwacje w systemie są rejestrowane (np. logi systemów, dane z czujników, interakcje użytkowników). Przykładowo platforma *glitchlab* generuje bogate logi kontekstu i wyników w procesie generowania kodu przez AI [27].
14. **Metryki i walidacja jakości** – pipeline powinien na bieżąco liczyć **metryki jakości** danych i działania AI. W glitchlab zdefiniowano np. metryki funkcjonalne (czy wynik działa) oraz metryki jakości struktury (czy wynik jest czytelny, zgodny ze stylem), a następnie wprowadzono **reguły werdyktu** hierarchizujące te kryteria: najpierw funkcjonalność, potem jakość, na końcu koszt czasowy [20]. To **modeluje hierarchię wartości** i zapewnia, że słaby wynik jest odrzucany zanim zmarnuje czyjś czas – co obniża koszt błędnych iteracji.
15. **Bramki decyzyjne (gating)** – na kluczowych etapach pipeline'u instaluje się automatyczne lub półautomatyczne **gates**, które decydują "go/no-go" dla artefaktów danych lub modeli. Projekt *sbom* (DevSecOps dla bezpieczeństwa oprogramowania) pokazuje taki mechanizm: każde wygenerowanie listy zależności (SBOM) i wynik skanowania bezpieczeństwa przechodzi przez bramkę – jeśli ryzyko przekracza próg, pipeline zatrzymuje się lub wymaga akceptacji wyjątku [28]. Ważnym wzorcem jest tu **kalibracja progów** – np. w fazie pilotażowej dopuszcza się tryb

"warn-only" (bramka przepuszcza z ostrzeżeniem), by zbierać dane o ryzyku i nie blokować innowacji [28]. Później progi się zaostrza. Dodatkowo system przewiduje **ścieżki wyjątków** (allowlist, akceptacja ryzyka przez człowieka) i **zamyka pętlę do organizacji** (np. integracja z systemem ticketowym do śledzenia, czy ktoś zareagował na ostrzeżenie) [42].

16. **Monitoring i obserwowalność** – cały system powinien dostarczać **telemetrii** o swoim działaniu: wskaźniki użycia, czasy przetwarzania, częstości błędów, interwencji człowieka itp. Projekt *swarm* – architektura dla floty dronów – podkreśla, że w systemach cyber-fizycznych **pełna obserwowalność** (monitoring stanu, bezpieczeństwa, sieci) jest krytyczna, bo koszty awarii są bardzo wysokie [38]. Bez danych telemetrycznych trudno optymalizować kosztów operacji AI i wykrywać, gdzie np. brak danych powoduje ryzyko. Organizacje powinny więc inwestować w **instrumentację metryk** zarówno dla działania modelu AI, jak i dla działań człowieka w pętli. NIST w swoim AI Risk Management Framework zaleca mierzenie m.in. częstości nadpisywania decyzji AI (override rate), czasu reakcji i liczby eskalacji [43] – to pozwala znaleźć wąskie gardła i policzyć realny **koszt nadzoru**. W naszej metodyce takie metryki sprzęgamy z ekonomiką: jeśli np. okazuje się, że człowiek odrzuca 30% rekomendacji AI, to znaczy, że jakość danych/modelu jest niewystarczająca i ROI będzie niższe (bo dużo czasu idzie na poprawki).

Dopiero **połączenie tych elementów** – od analizy wartości, przez model finansowy, po architekturę wykonawczą – daje pełen obraz opłacalności zbioru danych AI. W kolejnych sekcjach uszczegóławiamy niektóre z tych kroków, zaczynając od systematycznej klasyfikacji danych, poprzez wzorce kosztów i wartości, aż po ramy wdrożeniowe i praktyczne przykłady zastosowania metodyki.

Klasyfikacja typów danych a opłacalność

Nie wszystkie zbiory danych są sobie równe z perspektywy ekonomicznej. Proponujemy metodykę klasyfikacji, która pozwala ocenić profil danego zbioru pod kątem wartości i kosztów. Klasyfikacja obejmuje kilka wymiarów:

- **Surowe vs. produkt danych** – Jak wspomniano, dane **surowe** (raw) to takie, które wymagają jeszcze znacznego nakładu pracy zanim przyniosą wartość (np. nieoczyszczone logi, nieoznakowane obrazy). **Produkt danych** to zbiór przygotowany do użycia: posiada strukturę, dokumentację, często wstępne analizy lub modele. Surowe dane mogą być tańsze do zdobycia, ale wymagają inwestycji w przetworzenie – co obniża ich natychmiastową wartość rynkową. Produkty danych są *premium* – kupujący płaci za to, że ktoś już wykonał pracę i zredukował dla niego koszty przygotowania. W praktyce, opłacalność rośnie, gdy **awansujemy w łańcuchu wartości danych**: zamiast sprzedawać same dane, oferujemy je z kontekstem i jakością. Przykładem może być różnica między surowymi nagraniami wideo a gotowym zbiorem danych do trenowania samochodów autonomicznych (wraz z etykietami obiektów, scenariuszami testowymi i datasheetem). Ten drugi będzie miał wyższą cenę za jednostkę danych.

- **Dane krytyczne bezpieczeństwowo (safety-critical) vs. niekrytyczne** – Dane safety-critical to takie, które są wykorzystywane w systemach wysokiego ryzyka, gdzie błąd może skutkować zagrożeniem dla życia, zdrowia lub poważnymi stratami. Np. dane z sensorów w samolocie, sygnały monitorujące stan reaktora, dane treningowe do systemu wspomagania decyzji w chirurgii – to zbiory, gdzie **jakość i wiarygodność są bezwzględnym priorytetem**. Ich wartość rynkowa może być wysoka (bo klienci – np. szpitale, linie lotnicze – zapłacą za bezpieczeństwo), ale **koszt ich pozyskania i certyfikacji także jest ogromny**. Często wymagają one zgodności z normami (lotniczymi, medycznymi), audytów, a także ciągłego nadzoru człowieka (HITL). Z kolei dane niekrytyczne (np. preferencje muzyczne użytkowników) mają większą tolerancję na błędy – ich wartość wynika raczej z wolumenu i statystyki. Opłacalność danych krytycznych będzie oceniana według rygorystycznych kryteriów (koszt błędu, wymagane **redundancje i walidacje**),

podczas gdy dane masowe niekrytyczne – według skali (czy zbiór jest wystarczająco duży, by np. poprawić rekomendacje e-commerce). Warto zauważyć, że **dane safety-critical** często są **danymi rzadkimi** – bo pochodzą z unikalnych środowisk, np. z **autonomii w środowisku wysokich konsekwencji** (zarządzanie zasobami w habitacie kosmicznym) [17]. Takie dane nie mają odpowiedników na Ziemi, co czyni je rynkowo cennymi dla treningu AI (np. AI do zarządzania awariami).

- **Dane biologiczne i behawioralne vs. techniczne** – To podział ze względu na naturę zjawisk. **Dane biologiczne/behawioralne** dotyczą ludzi lub organizmów – np. dane medyczne pacjentów, dane o zachowaniu załogi w izolacji, pomiary biologiczne w ekosystemie zamkniętym. Charakteryzują się tym, że często są wrażliwe (kwestie prywatności), zmienne osobniczo oraz trudne do sztucznego wygenerowania (nie zasymulujemy łatwo realnych reakcji psychofizycznych na długotrwały pobyt w kosmosie). Ich pozyskanie bywa kosztowne (eksperymenty kliniczne, długotrwałe obserwacje) i musi uwzględniać etykę. Z drugiej strony, popyt na nie jest duży, jeśli wiążą się np. z poprawą zdrowia, wydajności pracy czy bezpieczeństwa. **Dane techniczne** (np. telemetryczne z urządzeń, dane produkcyjne z robotów, logi systemów IT) są zazwyczaj łatwiej standaryzowalne, generowane w dużych ilościach automatycznie. Bywają mniej wrażliwe (choć nie zawsze – np. logi mogą zawierać dane osobowe), ale też szybciej stają się *commodity* (bo wiele firm generuje podobne logi). Interesującą kategorią są **dane z zamkniętych systemów podtrzymania życia** – łączą aspekt techniczny i biologiczny. Np. dane z systemu ECLSS (Environmental Control and Life Support System) o obiegu wody, powietrza itp. w habitacie – to dane techniczne, ale krytyczne dla biologicznego przeżycia załogi. Europejski program **MELiSSA** (Micro-Ecological Life Support System Alternative) definiuje architekturę takich zamkniętych ekosystemów (bioreaktory, uprawy, filtracje) i ustanowił kryteria oceny ALISSE: masa, energia, efektywność, bezpieczeństwo, czas załogi [44]. Te kryteria można traktować jako zestaw wymiarów oceny danych z systemów bioregeneracyjnych – np. jak bardzo dane pomagają zmniejszyć masę uzupełnianych zasobów, ile energii kosztuje ich przetwarzanie, czy są bezpieczne dla załogi, ile uwagi ludzkiej wymagają. Dane biologiczne i techniczne będą inaczej wyceniane: pierwsze mogą dać przełomowe insighty (np. medycyna kosmiczna), drugie – optymalizacje procesów.

- **Dane do autonomii vs. dane do analiz (offline)** – **Dane do systemów autonomicznych** (np. do trenowania autopilota drona, systemu zarządzania habitatem) muszą często spełniać warunki czasu rzeczywistego, wysokiej niezawodności i kompletnie pokrywać przestrzeń przypadków, by AI nie „zgłupiała" w krytycznej sytuacji. Tutaj ważne są **dane skrajne, rzadkie przypadki** – bo to one często decydują o bezpieczeństwie. Z kolei **dane do analiz offline** (np. big data do wyciągania trendów, analizy biznesowej) mają inną dynamikę – mogą być przetwarzane hurtowo, kluczowa jest ich ilość i reprezentatywność statystyczna. Te drugie podlegają prawom ekonomii skali: im więcej danych, tym lepsze analizy do pewnego momentu, potem zwrot maleje. Dane do autonomii są bardziej „mission-critical" – tu każda dodatkowa sytuacja brzegowa uchwycona w danych może drastycznie zwiększyć wartość (bo zapobiegnie wypadkowi), ale trudno oszacować z góry ich wystarczalność. W ocenie opłacalności należy więc inne wagi przypisać *coverage* danych: dla autonomii – maksymalne pokrycie przypadków (nawet przy dużym koszcie), dla analiz – optymalna próba vs koszt.

- **Dane surowe vs. wzbogacone przez AI (syntetyczne)** – Nowym trendem jest użycie AI do wzbogacania danych lub generowania danych syntetycznych. Trzeba zadać pytanie: czy nasz zbiór zawiera *tylko dane rzeczywiste*, czy też jest uzupełniony danymi symulowanymi, augmentowany etc. Dane syntetyczne mogą obniżyć koszt pozyskania (łatwiej generować scenariusze niż je zebrać), ale ich wartość bywa niższa jeśli nie odzwierciedlają idealnie realnego świata. Istnieją przypadki, gdzie dane syntetyczne osiągają wysoki poziom opłacalności – np.

symulacje do szkolenia AI w sytuacjach niebezpiecznych (bo tańsze niż inscenizowanie prawdziwych katastrof). W klasyfikacji warto zaznaczyć, czy źródłem danych jest **realne środowisko** czy **symulator/algorytm**, bo to wpływa na zaufanie odbiorców (w niektórych branżach nie zaakceptują danych syntetycznych jako podstawy decyzji) oraz na koszty (generowanie może mieć swój koszt obliczeniowy i wymagać walidacji).

Podsumowując, klasyfikacja danych według powyższych kryteriów pozwala stworzyć **profil opłacalności** danego zbioru. Na przykład: *"produkt danych safety-critical z domeny biologicznej, generowany częściowo syntetycznie dla trenowania autonomicznego systemu medycznego"* – to opis, który sugeruje wysokie koszty (bo safety-critical, biologiczne), wysoką potencjalną wartość (unikalne, ratujące życie dane do AI), konieczność silnego governance (dane medyczne, prywatność) i mały margines błędu. Taki profil będzie oceniany inaczej niż np. *"surowe dane techniczne z czujników przemysłowych do analiz predykcyjnych maintenance"* – które są bardziej masowe, mniej wrażliwe i raczej nastawione na oszczędności kosztowych niż unikalną przewagę. **Metodyka architektoniczna wykorzystuje tę klasyfikację jako pierwszy krok**: pozwala szybko zidentyfikować największe wyzwania i dźwignie wartości dla danego zbioru.

Wzorce kosztów i wartości danych

Na podstawie powyższej klasyfikacji można przejść do **analizy wzorców kosztowych i wartości** dla różnych typów danych. Poniżej omawiamy kilka kluczowych wzorców ujawnionych w dostarczonych raportach i na rynku, ilustrując je konkretnymi danymi liczbowymi.

- **Koszty stałe vs. korzyści skali**: Inwestycje wymagające wysokiego CAPEX (np. infrastruktura kosmiczna, specjalistyczne laboratoria) muszą być zrównoważone przez odpowiednio duże przychody lub oszczędności. Jak wspomniano, ISS kosztuje ~4 mld USD rocznie [4], a planowana stacja komercyjna Starlab ~3 mld USD jednorazowo [3]. Tego rzędu wielkości nadają ton dyskusji o opłacalności – **próg rentowności** w takich projektach to setki milionów USD rocznie wygenerowanej wartości [31] [45]. Wiemy jednak, że same **przychody ze sprzedaży danych** rzadko osiągają miliard rocznie. Wspomniane firmy Earth Observation to maksymalnie ćwierć miliarda USD/rok [23]. Dlatego wyłania się wzorzec: **model czysto data-only nie pokryje olbrzymich kosztów stałych**, chyba że dane są naprawdę bezcenne lub… model biznesowy zostanie rozszerzony (hybrydyzacja). Hybrydą może być np. połączenie sprzedaży danych z **usługami eksperymentalnymi** (udostępnianie infrastruktury jako usługi – analogicznie do ISS National Lab) lub z **kontraktami instytucjonalnymi** zapewniającymi bazowe finansowanie [35] [36]. W raportach pokazano, że **ISS National Lab** działa właśnie jako taka hybryda: NASA finansuje bazę (~15 mln USD rocznie na zarządzanie, łącznie niemal 200 mln w wieloletniej perspektywie) [46], a sektor prywatny wnosi eksperymenty (w FY24 ponad 100 ładunków, ~80% komercyjnych) [47]. Mamy więc publiczny anchor plus komercyjny popyt – co razem tworzy działający ekosystem R&D na orbicie. Dla naszego modelu oceny znaczy to, że **wartość danych** czasem nie występuje samodzielnie, lecz w parze z wartością infrastruktury/usługi. W ocenie ROI takiego projektu trzeba uwzględnić dwa składniki: ROI z danych + ROI z usług (np. wynajmu platformy badawczej).

- **Wartość danych rzadkich i efekt nasycenia**: Dane **unikalne** potrafią osiągać astronomiczne ceny początkowe, ale tylko dopóki są unikalne. W jednym z raportów padł przykład, że prywatny rynek zapłaciłby (hipotetycznie) za „brakujące dane habitatu" nawet >1 mld USD rocznie, jeśli rzeczywiście rozwiązywałyby krytyczne problemy [48]. Jednak to założenie sprawdzono przez pryzmat realnych modeli – i okazało się, że *takie sumy się nie materializują bez udziału instytucji*. Największe realne kontrakty na dane+AI to ~290 mln USD/5 lat (przykład: kontrakt Maxar z

agencją wywiadowczą NGA) [49], co znów wskazuje na obecność **anchora instytucjonalnego**. Wartość rzadkich danych jest więc wysoka, ale **rynek prywatny ma ograniczoną chłonność**, chyba że dane są bezpośrednio powiązane z generowaniem zysku dla kupującego (np. dane finansowe dające przewagę na giełdzie – tam zapłacą dużo). Gdy oceniamy opłacalność zbioru, musimy prognozować, jak długo pozostanie on rzadki. Jeśli np. jesteśmy pionierem (first mover) w pozyskaniu pewnego rodzaju danych biologicznych z lotów kosmicznych, to przez pewien czas możemy dyktować cenę. Ale **„wysoka cena nie jest wieczna"** – inni wejdą na rynek, wzrosną wolumeny i cena spadnie do kosztu krańcowego pozyskania [11]. Dlatego metodyka powinna obejmować plan **dywersyfikacji lub upgrade'u wartości**: co dalej, gdy dana klasa informacji się zbanalizuje? Czy mamy kolejne innowacyjne zbiory w pipeline'ie? Czy zbudowaliśmy wokół danych ekosystem usług (np. narzędzia analityczne, benchmarking suite), który utrzyma klientów nawet jak same dane staną się tańsze? Trend rynkowy jest taki, że firmy data-driven starają się **przechodzić „w górę" – od sprzedaży surowych danych do sprzedaży insightów, modeli, rozwiązań AI opartych o te dane** [50]. Np. Planet Labs zaczynało od zdjęć satelitarnych, a obecnie oferuje **„AI-enabled solutions"** nad tymi danymi [50]. To podnosi średnią wartość kontraktu i buduje relacje długoterminowe, co stabilizuje przychody.

- **Koszt błędu i wartość marginalna danych**: W systemach, gdzie dominującym kosztem jest ryzyko błędu, wartość danych objawia się w sposób nie wprost – jako oszczędności przyszłych strat. Tutaj szczególnie przydaje się analiza **EV(ryzyka)**. Przykładowo, jeśli błąd diagnostyczny w medycynie kosztuje średnio 500 tys. USD odszkodowania, a obecny model AI ma 1% ryzyka takiego błędu na pacjenta, to dla 1000 pacjentów oczekiwany koszt błędów to 5 mln USD. Jeśli nowy zbiór danych obniży ryzyko do 0,5%, to oczekiwany koszt spada do 2,5 mln – oszczędność 2,5 mln USD. To nasz **ΔEV**, który jest bezpośrednią ekonomiczną wartością danych. Oczywiście, osiągnięcie tej redukcji może wymagać np. dodatkowego lekarza w pętli (co kosztuje) – więc musimy porównać ΔEV z ΔC (dodatkowym kosztem nadzoru) [51]. W badaniach nad HITL wykazano, że **HITL opłaca się, gdy koszt błędu jest duży, a koszt dodatkowej kontroli relatywnie mały** [14]. Dlatego w naszej metodyce oceniamy, czy dla danego zbioru istnieją **tanie metody kontroli jakości** (automatyczne bramki, sampling, interfejs usprawniający weryfikację) [14] – jeśli tak, łatwiej spełnić nierówność ROI > 0. Wartość marginalna danych przejawia się też w tzw. **ściętej granicy kompetencji AI** – AI dobrze radzi sobie z typowymi przypadkami, ale na brzegach (edge cases) nadal potrzebuje albo lepszych danych, albo człowieka. Dokładając nowe dane, najpierw eliminujemy duże błędy (duży spadek EV), potem zostają coraz rzadsze błędy (coraz mniejszy spadek EV). W pewnym momencie może się okazać, że koszt dodania kolejnych danych (np. wykonania kolejnych 100 eksperymentów) przewyższa zysk z redukcji już bardzo małego ryzyka. **Optimum ekonomiczne** jest tam, gdzie krańcowy koszt danych = krańcowa korzyść z redukcji ryzyka. Nasza metodyka, poprzez iteracyjne prototypowanie, stara się ten punkt uchwycić – żeby nie „przeinwestować" w dane dla perfekcji, która się finansowo nie zwróci.

- **Autonomia vs. udział człowieka – efekt na OPEX**: W projektowaniu systemu generowania danych (np. habitat zbierający dane operacyjne o sobie) pojawia się pytanie o poziom autonomii. Paradoksalnie, aby dane były wartościowe (np. pokazywały unikalne zjawiska), nieraz potrzeba udziału ludzi – choćby do tego, by stworzyć sytuacje, które generują interesujące dane (np. ludzka współpraca w izolacji). Jednak ludzie generują ogromny koszt OPEX (załoga, jej utrzymanie, życie). Zatem jednym z wzorców kosztowych jest **inwestycja w autonomię w celu redukcji OPEX**, ale tak, by nie stracić na wartości danych. W raporcie o pięciu modelach wioski kosmicznej wskazano, że kluczowe mechanizmy poprawy ekonomii to **autonomia + „ekologia" (zamykanie pętli)** – one redukują koszty operacyjne [52]. Autonomia (AI-ops) zmniejsza potrzebę pracy ludzkiej, a zamykanie pętli obiegu (recykling zasobów) zmniejsza koszty zaopatrzenia. Jednak uwaga: zbyt pełna autonomia może **zubożyć dane** (bo mniej interakcji człowieka = mniej danych behawioralnych, które były cenne), a pełna ekologia jednego

czynnika (np. wody odzyskiwanej w 98% [53]) nie oznacza, że inne koszty znikną (żywność, części zamienne pozostają dużym ciężarem [54]). Wartość danych może nawet spaść, jeśli system jest zbyt idealnie autonomiczny – bo nie generuje „case'ów" pokazujących jak człowiek sobie radzi. Stąd opłacalny bywa **model hybrydowy**: część procesów silnie autonomiczna (redukcja wydatków), ale pewne elementy pozostają manualne/kontrolowane przez ludzi, by dostarczać dane o zachowaniach, decyzjach, które są rzadkie i cenne. Przykładowo, **model „rotacje Social-AI"** zakłada rotacyjne misje ludzi współpracujących z AI – generuje to dane o interakcji człowiek-AI, a jednocześnie nie wymaga stałej obecności dużej załogi (ograniczenie OPEX) [55] [56]. Ogólnie, wzorzec jest taki: maksymalizować dane per koszt załogi. Jeśli 1 astronauta generuje tyle danych co dawniej 3 astronautów dzięki dobrym narzędziom AI, to mamy lepszą opłacalność.

- **Dwustronna platforma danych**: Pojęcie **platformy dwustronnej** oznacza model biznesowy, gdzie jest dostawca danych i użytkownik danych spotykający się na platformie (np. marketplace). Aby platforma prosperowała, zazwyczaj jedna strona jest subsydiowana, by przyciągnąć drugą (klasyczny przykład: kierowcy i pasażerowie w Uberze – początkowo dopłaty dla kierowców, zniżki dla pasażerów). W danych może to oznaczać, że np. udostępniamy pewne dane darmowo lub poniżej kosztu, by zbudować rynek usług na tych danych. W kontekście habitatów kosmicznych analiza pokazała, że **przynajmniej jeden strumień musi być „fundamentem stałych kosztów"** – np. stały kontrakt z agencją (anchor), abonament od konsorcjum firm – bo trudno monetyzować jednocześnie wszystkich na wolnym rynku bez takiej podstawy [57] [34]. Wartość danych w takim modelu może być więc częściowo realizowana w modelu freemium: dajemy trochę danych by zachęcić (generując ukryty zysk np. w postaci standardów czy efektu sieciowego), a zarabiamy na premium dostępie lub na usługach analitycznych. Ekonomicznie oceniając zbiór danych, powinniśmy spojrzeć, czy może on stać się elementem większej platformy – jeśli tak, to jego wartość rośnie (może generować przychody pośrednie, zwiększać lojalność klientów, napędzać sprzedaż innych produktów).

Każdy z powyższych wzorców powinien zostać wzięty pod uwagę w **ramach oceny opłacalności**. Metodyka architektoniczna zakłada, że zamiast traktować projekt danych statycznie, analizujemy jego **mechanikę ekonomiczną** – tzn. jakie dźwignie kosztów i wartości można zastosować. Przykładowo, jeżeli widzimy, że projekt balansuje na granicy opłacalności, bo koszty nadzoru jakości są wysokie – można w warstwie architektury zaproponować więcej automatyzacji walidacji (inwestycja w narzędzia QA, jak w *sbom*) lub zmianę workflow (np. wzorzec triage: **przekazać AI prostsze 80% przypadków, a ekspertom tylko 20% najtrudniejszych** [58]). Takie heurystyki potrafią poprawić ROI projektu bez zmiany samego zbioru danych – ale przez zmianę sposobu jego wykorzystania. Dlatego pełna ocena opłacalności zawsze będzie iteracyjna: od charakterystyki danych -> przez model finansowy -> po propozycje usprawnień architektonicznych, które z powrotem wpływają na koszty i wartości.

Ramy wdrożeniowe metodyki i przykłady zastosowań

Proponowana metodyka musi znaleźć odzwierciedlenie w konkretnych procesach wdrożeniowych. Poniżej przedstawiamy **ramy organizacyjne** i przykłady, jak tę metodologię zastosować w praktyce – od ziemskiego inkubatora innowacji danych, przez orbitalny habitat, po koncepcję „taśmy prototypowej" szybkiego wytwarzania danych.

1. Wdrożenie w inkubatorze danych AI (projekt poniżej progu startupu) – Środowisko inkubatora lub akceleratora startupów jest idealne do zastosowania naszej metodyki w skali mikro. Tutaj zwykle dysponujemy ograniczonym budżetem i czasem, a celem jest zweryfikować pomysł na wykorzystanie danych zanim powstanie pełnoprawna firma. Ramy metodyki sugerują: - **Małe, zwinne zespoły**

projektowe: Zespół 2-5 osób o uzupełniających się kompetencjach (data engineering, domain expert, biznes). Stosujemy model „mrowiska" – role elastyczne, każdy wnosi wiele ról, aby zmniejszyć koszty osobowe [39]. - **Krótki cykl prototypowania**: Zamiast budować od razu kompletny produkt, tworzymy iteracyjnie prototypy danych. To co raport określa mianem **„taśmy prototypowej danych" poniżej progu startupu** – czyli **system produkcyjny do szybkiego wybierania i testowania „braków danych"** [59] [60]. W praktyce inkubator może np. co 3-6 miesięcy uruchomić mini-projekt danych: wybieramy jeden high-value *data gap*, tworzymy MVP-Data, sprawdzamy zainteresowanie rynku. - **Finansowanie etapowe (mix grantów i inwestorów)**: Wzorem **programów SBIR** NASA – faza I mała (<$150k, 6 mies.) na demonstrację pomysłu, faza II większa (<$850k, 24 mies.) na rozwój [61]. Inkubator może zapewnić grant zalążkowy (np. 50 000 EUR w programie ESA BIC) oraz dostęp do infrastruktury i mentorów [62]. To pokrywa CAPEX pierwszego rzędu (sprzęt, dane referencyjne, praca zespołu). Jeśli faza I pokaże **dowody popytu i unikalności danych**, można się starać o kolejne finansowanie (VC, kontrakt pilotażowy z klientem). Ważne jest ustalenie **„kill criteria"** – jasno określonych warunków, kiedy projekt zamykamy, jeśli nie rokuje (np. brak poprawy metryk lub brak zainteresowania w ciągu X miesięcy) [63]. Taka dyscyplina chroni przed topieniem kosztów. - **Mentoring i dostęp do środowisk testowych**: Inkubator powinien ułatwić dostęp do unikalnych zasobów potrzebnych do zebrania danych – np. loty testowe suborbitalne w programie NASA Flight Opportunities [64] lub miejsca na ładunek w ISS National Lab (który uruchomił program **Orbital Edge Accelerator** łączący inwestycje 500k USD ze sponsorowanym dostępem do eksperymentu na orbicie) [65]. W Europie ESA BIC wraz z ośrodkami badawczymi (np. NCBJ w Polsce) pełnią podobną rolę – zapewniając zaplecze technologiczne do prototypów [62] [66]. Przykładem praktycznym może być startup pracujący nad danymi z mikrograwitacji: inkubator pomaga mu wysłać eksperyment na ISS, by wygenerować pierwsze **unikalne dane** i sprawdzić, czy ktoś (np. firma farmaceutyczna) zechce za nie zapłacić. Takie podejście minimalizuje ryzyko – zamiast budować od razu drogi habitat kosmiczny, testujemy koncept małymi krokami. - **Pętla feedbacku rynku**: Każdy prototyp danych powinien trafić do potencjalnych użytkowników jak najszybciej – np. w formie pilotażu, konkursu (hackathon z użyciem tych danych) lub po prostu rozmów z klientami, gdzie pokazujemy próbkę danych i pytamy o model biznesowy. To realizacja zasady, że *popyt płaci za rozwiązanie problemu, a nie za dane same w sobie* [18]. Jeśli feedback jest negatywny („dane nie rozwiązują naszego problemu"), projekt jest pivotowany lub zamykany. Jeśli pozytywny – inkubator pomógł zweryfikować opłacalność i można skalować.

2. Zastosowanie w habitatcie orbitalnym (wioska kosmiczna) – Rozważmy teraz na drugim skraju skali scenariusz dużego projektu: zamieszkały habitat na orbicie, który ma się utrzymać ekonomicznie z **produkcji danych dla AI**. Tutaj metodyka musi objąć bardzo złożony system, ale zasady pozostają te same: - **Synergiczny ekosystem ludzi i AI**: Habitat to środowisko Social-AI, gdzie ludzie i AI współdziałają, by jednocześnie utrzymać się przy życiu i generować wartościowe dane [67]. Nasza metodyka klasyfikuje trzy główne klasy danych z takiego habitatu: **(1)** dane operacyjno-decyzyjne z autonomicznego zarządzania w warunkach high-stakes (awarie, alokacja zasobów) – to dane typu safety-critical/autonomy, **(2)** dane biologiczno-behawioralne o ludziach w izolacji i nieważkości (zdrowie, adaptacja, zachowania społeczne) – unikalne dane human factors, **(3)** dane z systemów zamkniętej pętli (wydajność recyklingu, stabilność ekosystemu) [17]. Te wszystkie są *danymi brakującymi* na rynku, więc potencjalnie cennymi. Architektura habitatu powinna być projektowana tak, by **maksymalizować zbieranie tych danych** bez zakłócania podstawowej misji. Np. infrastruktura sensoryczna od początku musi służyć podwójnie – do operacji i do logowania danych dla analityki. - **Model biznesowy hybrydowy**: Z poprzednich analiz wiemy, że habitat nie utrzyma się wyłącznie ze sprzedaży surowych datasetów kosmicznych. Dlatego wdrażamy model hybrydowy: poza sprzedażą danych (np. licencje na unikalne zbiory dot. zachowania załogi) mamy **kontrakty anchor** – np. umowę z agencją rządową na zapewnienie platformy R&D (analog ISS NL, ale komercyjny) – oraz usługi **EaaS (Environment-as-a-Service)** dla firm, które chcą przeprowadzić eksperymenty w mikrogrze (np. testy materiałów, farmacja) [68] [69]. W ten sposób co najmniej jeden strumień przychodów jest pewny i długoterminowy (np. agencja płaci stałą kwotę rocznie za gotowość platformy), a dane stają się *produktem ubocznym* tych

działań. Oczywiście, dążymy, aby dane stały się z czasem produktem głównym – ale realia uczą, że potrzeba takiego kotwicznego finansowania. Nasza metodyka w fazie planowania finansowego habitatu kładzie nacisk na znalezienie **NPV >= 0** przy realistycznych założeniach: jeśli czysta sprzedaż danych tego nie daje, to uwzględniamy inne usługi lub subsydia. Test warunku rentowności jest tu bezlitosny: pokazuje „twardy próg sprzedaży" niezależny od narracji [70] – np. jeśli wyliczymy, że trzeba 1,5 mld USD/rok przychodu, a rynek danych wskazuje max 0,2 mld, wiemy że musimy zmienić założenia. - **Autonomia i załoga – optymalny miks**: Zastosowanie modelu z architekturą HITL i ant-kolonią (mrowiskiem) pozwala zminimalizować koszty załogi, nie tracąc danych. Założymy np., że habitat operuje nominalnie z 4 osobami załogi (dla bezpieczeństwa i obsługi), a resztę czynności wykonuje AI (roboty, automatyka). W sytuacjach szczególnych (np. misje badawcze) załoga może wzrosnąć rotacyjnie do 10 osób, co generuje skok danych (więcej interakcji człowiek-człowiek i człowiek-AI), ale na ograniczony czas [55]. W ten sposób **zarządzamy OPEX**: stały OPEX niższy dzięki autonomii, okresowe wyższe koszty tylko gdy potrzebne dla generowania nowej porcji danych premium. Ponadto, zastosujemy maksymalnie zamknięte systemy (zgodnie z MELiSSA) by zmniejszyć koszty logistyczne – co raport określa jako *ekologia jako dźwignia kosztowa* [71]. Pamiętajmy jednak o kontrapunkcie: nawet 98% odzysk wody nie eliminuje wszystkich dostaw, a rzeczy takie jak żywność i części zamienne nadal generują duże koszty [54]. Stąd plan finansowy musi zakładać pewien minimalny **strumień kosztów stałych**, którego nie da się „innowacyjnie" pozbyć. - **Zarządzanie danymi i jakością w locie**: Habitat to organizm, w którym nie ma miejsca na błędy – więc jakość danych jest też kwestią bezpieczeństwa. Wdrożymy standardy dokumentacyjne jak datasheets i model cards dla wszelkich modeli operujących na danych habitatu (np. AI monitorującej zdrowie załogi) [72]. Każdy zbiór danych będzie miał swojego **„kustosza"** – osobę (lub AI) odpowiedzialną za jego monitoring i aktualność. Na poziomie organizacyjnym utworzymy role jak **inżynier ds. jakości i bezpieczeństwa AI** (przykład z dokumentu: AI Safety & Assurance), który dba o audyty modeli, testy i kryteria dopuszczenia modeli do zadań safety-critical [73] [74]. Będzie on korzystał z **model cards** i procedur walidacji zanim model zacznie działać autonomicznie. Taki governance to dodatkowy koszt (element OPEX), ale niezbędny – incydent z AI na stacji może mieć katastrofalne skutki finansowe i ludzkie, więc lepiej zapobiegać (koszt kontroli) niż leczyć (koszt błędu). - **Benchmarking i wartość rynkowa danych z habitatu**: By zmaksymalizować wartość rynkową danych z orbitującej „wioski", planujemy aktywnie włączać je w szerszy ekosystem naukowo-biznesowy. Np. dane z kategorii (1) – autonomicznego zarządzania – można wykorzystać do stworzenia **benchmarku dla algorytmów zarządzania awariami**. Habitat publikuje (odpłatnie lub we współpracy) część danych jako wyzwanie dla zespołów AI: kto stworzy lepszy algorytm zarządzania kryzysowego na bazie tych danych. To przyciąga uwagę i może wykreować standard de facto (jeśli Twój algorytm działa na danych z ISS/habitatu, to znaczy że jest dobry). Podobnie dane biologiczne mogą zasilić badania medycyny kosmicznej, co przerodzi się w granty i współpracę z uniwersytetami lub firmami biotech. Tutaj **wartość danych wyrażana jest nie tylko w sprzedaży bezpośredniej, ale w ich** wartości sieciowej **– stają się platformą współpracy R&D. Ekonomicznie można to traktować jako zwiększanie** wartości przyszłych opcji** (real options): nawet jeśli teraz nie monetyzujemy każdej próbki, to budujemy kapitał intelektualny i relacje, które zaowocują np. kontraktem na wspólne prace B+R (który zasili budżet habitatu).

Podsumowując przykład habitatu: nasza metodyka pozwala już na etapie projektu odpowiedzieć, **czy habitat „data-only" ma szanse być opłacalny** (wygląda na to, że nie bez hybrydowych przychodów) oraz **jakie warunki muszą być spełnione, by w ogóle zbliżyć się do rentowności** (wysoki anchor tenant, dywersyfikacja strumieni, ciągła innowacja w danych rzadkich, automatyzacja dla kosztów). To cenny wgląd dla decydentów – chroni przed inwestycją w „kosmiczny data center", który by się nie zwrócił. Co ważne, wskazuje też **gdzie jest wartość**: w danych unikalnych zanim staną się powszechne, w kontraktach długoterminowych, w obniżaniu kosztów poprzez AI i recykling, oraz w integracji człowieka w pętli tam, gdzie daje to dodatkowe bezpieczeństwo i dane.

3. „Taśma prototypowa danych" – seryjne przekuwanie braków danych w wartość – Ten koncept przewijał się już powyżej, ale warto go na koniec uwypuklić jako ogólny framework wdrożeniowy. **Taśma prototypowa** to inaczej **proces innowacji w obszarze danych**, który pozwala organizacji systematycznie tworzyć nowe zbiory danych o wysokim ROI: - Działa to jak **taśma produkcyjna**, ale zamiast produktów fizycznych – prototypy danych. Każdy cykl zaczyna się od zidentyfikowania *luki danych* na rynku, potem szybkie złożenie aparatury/zdobycie źródła, zebranie próbki danych, walidacja użyteczności (np. mały eksperyment AI), zapakowanie w produkt (dokumentacja, interfejs API, może wstępny model), i decyzja: czy inkubujemy to dalej w startup/projekt, czy zamykamy. - Kluczowe atrybuty taśmy: **niski koszt wejścia, krótki czas cyklu, możliwość przerwania na każdym etapie** [60]. To jest przeciwieństwo wielkich projektów „najpierw zbudujmy całą infrastrukturę, a potem zobaczymy dane" – tutaj minimalna infrastruktura (często korzystamy z istniejących platform – np. wysyłamy eksperyment na balonie zamiast budować satelitę; albo korzystamy z analogów misji kosmicznych na Ziemi, żeby symulować dane z Marsa). NASA praktykuje taki **phased approach** od lat: np. zanim zainwestuje się w drogi instrument satelitarny, testuje się prototyp na rakiecie suborbitalnej (Sounding Rockets jako *low-cost testbed*) [75]. Z punktu widzenia naszej metodyki, taśma prototypowa to **strategia zarządzania ryzykiem**: każdy kolejny etap jest opcją, z której korzystamy tylko, gdy poprzedni udowodnił wartość [60]. - Organizacyjnie, wdrożenie takiej taśmy wymaga pewnej kultury i struktur. Firma/instytucja musi mieć wydzielony zespół lub jednostkę R&D, która ma mandat na eksperymenty i porażki (bo wiele prototypów zostanie „ubitych" – i to jest sukces, nie porażka, bo oszczędziliśmy dalszych kosztów). Przydatne są **ramy formalne** jak wspomniane TRL (Technology Readiness Levels) – można je adaptować do oceny dojrzałości zbioru danych i modelu AI [76]. Np. TRL 1-3: koncept i dowód słuszności (mamy pomysł na zbiór danych i drobny eksperyment potwierdzający, że może być przydatny), TRL 4-6: prototyp w warunkach zbliżonych do docelowych (np. zbieramy dane na analogach lub na małej próbce w środowisku docelowym), TRL 7-9: pełna demonstracja i produkt (działający system zbierania danych na docelowej platformie, gotowy do skali). Dzięki TRL mamy **bramki decyzyjne**: nie przechodzisz do TRL 5, jeśli na TRL 3 dane nie wykazały wartości – co wpisuje się w naszą filozofię kill criteria. - Praktycznym wzorcem jest też łączenie **kapitału z dostępem do środowiska**. ISS National Lab swoim akceleratorem orbitalnym de facto tworzy taką taśmę – daje trochę pieniędzy i bilet na orbitę, byś mógł wygenerować *pierwsze kosmiczne dane* i z nimi wrócić na Ziemię do inwestorów [65]. W skali krajowej, ESA BIC wraz z instytutami badawczymi robi to samo – daje fundusze pre-seed i zaplecze labowe, aby startup mógł przez rok zbierać np. dane w laboratorium zamiast od razu budować fabrykę [62]. - Zaletą taśmy prototypowej jest **udowodnienie opłacalności bez czekania wielu lat**. Można np. zamiast budować habitat za miliardy, zrobić analog habitatowy na Ziemi (lub wykorzystać moduł ISS na parę miesięcy) i tam zebrać namiastkę danych (powiedzmy 3-miesięczny eksperyment z 4 osobami, generujący dane biologiczne i operacyjne). Następnie policzyć, ile warte były te dane: czy np. ktoś zapłacił za raport z tego eksperymentu, czy powstał model AI, który udało się sprzedać? Jeśli nie, to projekt kosmicznej wioski w wersji data-only raczej nie ma sensu – co stanowi **falsyfikację ekonomiczną** hipotezy przed dużym wydatkiem [77]. Tę właśnie filozofię przyjęto w cytowanych badaniach: zanim zrobimy „idealną wioskę kosmiczną", sprawdzono twarde liczby rynkowe i wyszło, że bez anchorów i hybrydy się nie domknie.

Na zakończenie, warto podkreślić, że przedstawiona **metodyka architektoniczna** jest **narzędziem uniwersalnym** – można ją skalować w górę (dla dużych przedsięwzięć jak orbitalne habitaty) i w dół (dla małych startupów danych). Łączy ona **perspektywę techniczną** (architektura zbierania i przetwarzania danych, standardy, pętle HITL), **perspektywę ekonomiczną** (ROI, NPV, koszty vs zyski, rynek) oraz **perspektywę organizacyjną** (role, procesy, modele współpracy). Dzięki temu pozwala ocenić i zaprojektować inicjatywy danych tak, by **maksymalizować ich wartość rynkową przy kontrolowanych kosztach i ryzyku**. W szybko zmieniającym się krajobrazie AI (oraz np. eksploracji kosmosu przez sektor prywatny) taka metodyka może być przewagą – umożliwia świadome podejmowanie decyzji, które z potencjalnych zbiorów danych rozwijać i jak to robić, by inwestycja się opłaciła. W erze, gdzie „dane to nowe paliwo", równie ważne jest by umieć ocenić **kaloryczność** każdego

litra tego paliwa zanim zainwestujemy w jego produkcję – i temu właśnie służy powyższy raport i zawarte w nim ramy postępowania.

Źródła: W opracowaniu wykorzystano analizy i dane liczbowe z dostarczonych dokumentów, m.in. raportów nt. ekonomii „wioski kosmicznej” [4] [3] [23], koncepcji „taśmy prototypowej danych” [60] [61], efektywności systemów Human-AI-In-the-Loop [40] [14] oraz wzorców projektowych glitchlab/sbom/swarm [27] [28] [38]. Przytoczone liczby (koszty, przychody) i przykłady wzorców pochodzą z tych źródeł dla uwiarygodnienia przedstawionej metodyki.

---

[1] [70] Ekonomiczny dowód opłacalności modelu „idealnej wioski kosmicznej” jako synergicznego ekosystemu lud.docx
file://file_00000000679871f4b34955fe888947ac

[2] [13] [14] [15] [16] [20] [27] [28] [29] [33] [37] [38] [40] [42] [43] [51] [58] Opłacalność ekonomiczna i kreacja wartości w systemach Human-AI In-The-Loop.docx
file://file_00000000614c720ab27f992dfbc94d0c

[3] [4] [5] [23] [26] [44] [52] [53] [54] [67] [71] Pięć racjonalnych modeli hybrydowej wioski kosmicznej Social-AI_ rachunek opłacalności, heurystyki s.docx
file://file_00000000294471f4a18036cb176cc2e0

[6] [39] [41] [73] [74] Kompetencje AI jako rdzeń „organicznej” wioski kosmicznej_ model mrowiska, intuicja na metapoziomie .pdf
file://file_0000000025e871f4808f72f4853298cc

[7] [8] [9] [25] [30] [31] [32] [34] [35] [36] [45] [55] [56] [57] [68] [69] [72] Pięć racjonalnych modeli hybrydowej wioski kosmicznej Social-AI_ rachunek opłacalności, heurystyki s.pdf
file://file_00000000cf10720ab70f5a43fe60c3d5

[10] [11] [12] [17] [21] [22] [24] [46] [48] [49] [50] [77] Ekonomiczna falsyfikacja modelu „data-only” dla wioski kosmicznej jako centrum produkcji danych dla .docx
file://file_00000000576871f48fd74901d818fd5c

[18] [19] [47] [59] [60] [61] [62] [63] [64] [65] [66] [75] [76] Ekonomiczna „taśma prototypowa danych” poniżej progu startupu_ jak seryjnie przekuwać brakujące dane.docx
file://file_0000000073a871f482ee98efbc60e136