

Sztuczna Inteligencja

Big CyC i supersystemy ekspertowe

Włodzisław Duch
Katedra Informatyki Stosowanej UMK
Google: Włodzisław Duch

Nadchodzi czas symbiozy

1. Sztuczna inteligencja: czym to jest?
Definicje, sytuacja obecna, przewidywania.
2. Jak się to ma do naszych mózgów?
3. AI: jak to wykorzystać.
Biurokracja, laborant, telemedycyna, alter ego,
roboty, wirtualne światy.
4. Czego można się spodziewać?
Naprawianie mózgów, nekromancja, uzależnienia ...

AI to znacznie więcej niż ChatGPT!
Widzimy tylko wierzchołek góry lodowej ...

Ponad 600 publikacji tygodniowo w repozytorium [arxiv cs.AI](https://arxiv.org/)



CYC



- Zdrowy rozsądek
- CYC - podstawy
- CYC - ontologia podstawowa
- CYC – możliwości i zastosowania

Zdrowy rozsądek



- Najbardziej ambitny regułowy system ekspertowy.
- Realizowany pod kierunkiem D. Lenata w MCI od 1984 roku, od 1995 w komercyjnej firmie Cycorp, www.cyc.com
- Zdrowy rozsądek wymaga milionów reguł! Czy to realne?
- Test ostateczny dla technologii ES - czy tak złożony system będzie działał?
- Początkowe oceny - 5% szans na użyteczność.

Problemy:

- „Kruchość” systemów ekspertowych:
- Niewiele systemów ma ponad 1000 reguł;
- którą regułę/ramę użyć? Kosztowne sprawdzanie warunków.
- Formy pytań musiały być ściśle określone, brak zrozumienia języka.

Zdrowy rozsądek cd



Ludzie nie dają się łatwo zbić z tropu.

- „Szeroka” wiedza $\Leftrightarrow$ odpowiedzi szybkie = rozumowania płytkie.
Potrzebujemy wielu sposobów kojarzenia faktów (reasoners).
- Baza wiedzy - fakty, heurystyki, wiedza specyficzna.
- System-partner człowieka?
- Musi znać nastawienia ludzi, światopogląd, mikroteorie zależne od kontekstu.

Problemy:

- Czy znane sposoby reprezentacji wiedzy wystarczą?
- Reprezentacja podstawowych pojęć - ontologia podstawowa.
- Czas, przestrzeń, ruch, substancja ... wspólna wiedza ES.
- Knowledge-based systems - podstawa porozumiewania się.

Zdrowy rozsądek cd.



- System CYC zawiera w wersji podstawowej ponad milion reguł.
- Baza wiedzy powstała w oparciu o analizy ekspertów.
- Bootstrapping - analiza tekstów w oparciu o zgromadzoną wiedzę miała pozwolić na automatyzację procesu akwizycji wiedzy – dopiero teraz to możliwe.
- Język reprezentacji wiedzy deklaratywnej + procedury manipulowania faktami.
- Język opisuje obiekty, zdarzenia, nastawienia.
- Język reprezentacji CycL rozwijał się wraz z KB.
- Oparty na ramach + rachunek predykatów + uzupełnianie zmiennych domyślnych.
- Nowe pojęcia mogą wymagać rozszerzenia CycL - „ekspresywność” języka nie jest dostateczna.
- Obecnie CycL jest stabilny.

The Cyc Advantage

LOGIC-BASED MACHINE REASONING



PERFORMANCE

Autonomous data gathering and logical decision making. Mitigates human error, avoidance of scrutiny and machine learning biases.



CONTEXT & UNDERSTANDING

Real-time and continuous semantically enriched representation for every order, for every patient, across every facility.



TRUST & AUDITABILITY

Real-time and continuous explainability and auditability to patient record.

Left
brain

$\forall x \forall y \forall z ((\text{owns } x y) \wedge (\text{parts } y z) \Rightarrow (\text{owns } x z))$
LOGIC
 P
 Q
 $P \wedge Q \Rightarrow R$

PATTERNS

Right
brain

One company provides broad logical/causal reasoning at enterprise scale: Cycorp

1000s of companies provide this type of statistical/Machine Learning AI

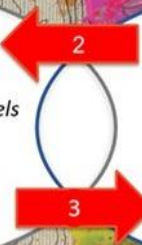
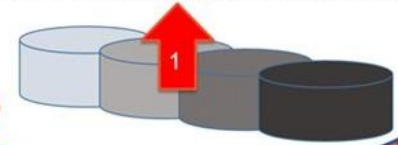
$A \rightarrow B \rightarrow C \rightarrow D \rightarrow E \dots \rightarrow Y \rightarrow Z$

Causation from Explicit Symbolic Logical Reasoning guided by heuristics and causal models



$A \rightarrow Z$

Correlations using big data: Statistical ML



CycL 1

Ramy - dziedziczenie dla specjalnych połączeń, *isa*, *instance*

- „Wszystkie ptaki mają dwie nogi”.
- Szufladka *nogi* w ramkach typu *ptak*, ma wartość „2”.
- „Wszyscy przyjaciele królika lubią marchewkę”.
- Ramki z *królik* w szufladce *przyjaciół* dziedziczą *marchewka* jako zawartość szufladki *lubi*.

Dziedziczenie przez łańcuch relacji:

- „wszyscy krewni przyjaciół królika mają długie uszy”.
- Szufladki *uszy* krewnych przyjaciół królika => *długie*.

CycL 2

Constraint language - opis przez specyfikację ograniczeń.

- „Norbert nie lubi ludzi, którzy programują w Fortranie”.
- Nie zapełnia szufladek stwierdzeniem: „Norbert ich lubi/nie lubi”
- Zmienna ograniczająca *fortran-constraint*, w szufladce „lubi” ramy Norbert.
- **Język specyfikacji ograniczeń** pozwala na realizację pełnej logiki pierwszego rzędu, uzupełnienie ram.
- Osiągnięto większą sprawność rozumowania opartego o ramy niż rozumowania wymagającego logicznych wniosków.

Cycl - mikroteorie

- Mechanizmy wnioskowania: uniwersalno-proceduralne.
- Identyfikacja często używanych klas wnioskowań
- => dodaje się sprawne mechanizmy.
- Specyficzne mechanizmy wnioskowania, zależnie od dziedziny.
- Wnioskowanie i tworzenie „mikroteorii”, np. opis posługiwania się pieniędzmi, robienie zakupów, jazda autem.
- 1. Pieniądze daje się w zamian za towary, rzeczy, usługi lub jako dar.
- 2. Każdy czynnik jest związany z pewną sumą pieniędzy.
- 3. Opłaty do 10\$ dokonywane są zwykle gotówką, powyżej 50\$ czekiem lub kartą kredytową.
-
- 17. Kiedy coś kupujesz jednym z podzdarzeń jest płacenie sprzedawcy lub w kasie.

Spójność wiedzy

CYC zawiera > 1100 różnych mechanizmów wnioskowania (reasoners, inference engines), łączących > 1.5 mln pojęć.

- Dla każdego z nich usiłuje zachować spójność wiedzy.
- Czy w szufladkach są dopuszczalne wartości ?

Jak rozstrzygać sprzeczności?

- Testy rozumienia: analiza opowiadań i pytania.
- Czas stosowania domyślnego rozumowania zależy od reguły.
- Reguły *if-needed rule* używane tylko gdy się do nich odwołujemy; reguły *if-added rule* oceniające wszystko, co się da.
- *Forward rules*, reguły wyprzedzające - dużo czasu, mało przydatne.
- Np. własności znajomego Jana „programuje w Fortranie” to nazwisko znajomego może się pojawić na liście osób, których Jan nie lubi.

Poziomy reprezentacji

- Podsystem utrzymywania spójności CYC (truth maintenance system). Dodaje nowe fakty w nocy lub wolnego czasu.
- Poziom epistemologiczny (EL) - abstrakcyjny poziom reprezentacji, wykorzystuje język specyfikacji ograniczeń, wymaga abstrakcyjnego rozumowania logicznego.
- Poziom reprezentacji heurystycznej (HL) - deklarowane fakty, efektywne heurystyczne sposoby rozumowania.
- Zbiór mechanizmów wnioskowania skojarzony jest z każdą szufladką. Pytanie proste - sprawdź zawartości szufladki.
Pytanie złożone - wnioski w celu zapełnienia szufladki.
- Strategia kontrolna wnioskowania - również w oparciu o ramy.
- Schematy wnioskowania w ramach, kopie zmodyfikowane stosuje się do nowych przypadków
- Kompilator języka CycL tworzy kod w Lispie.

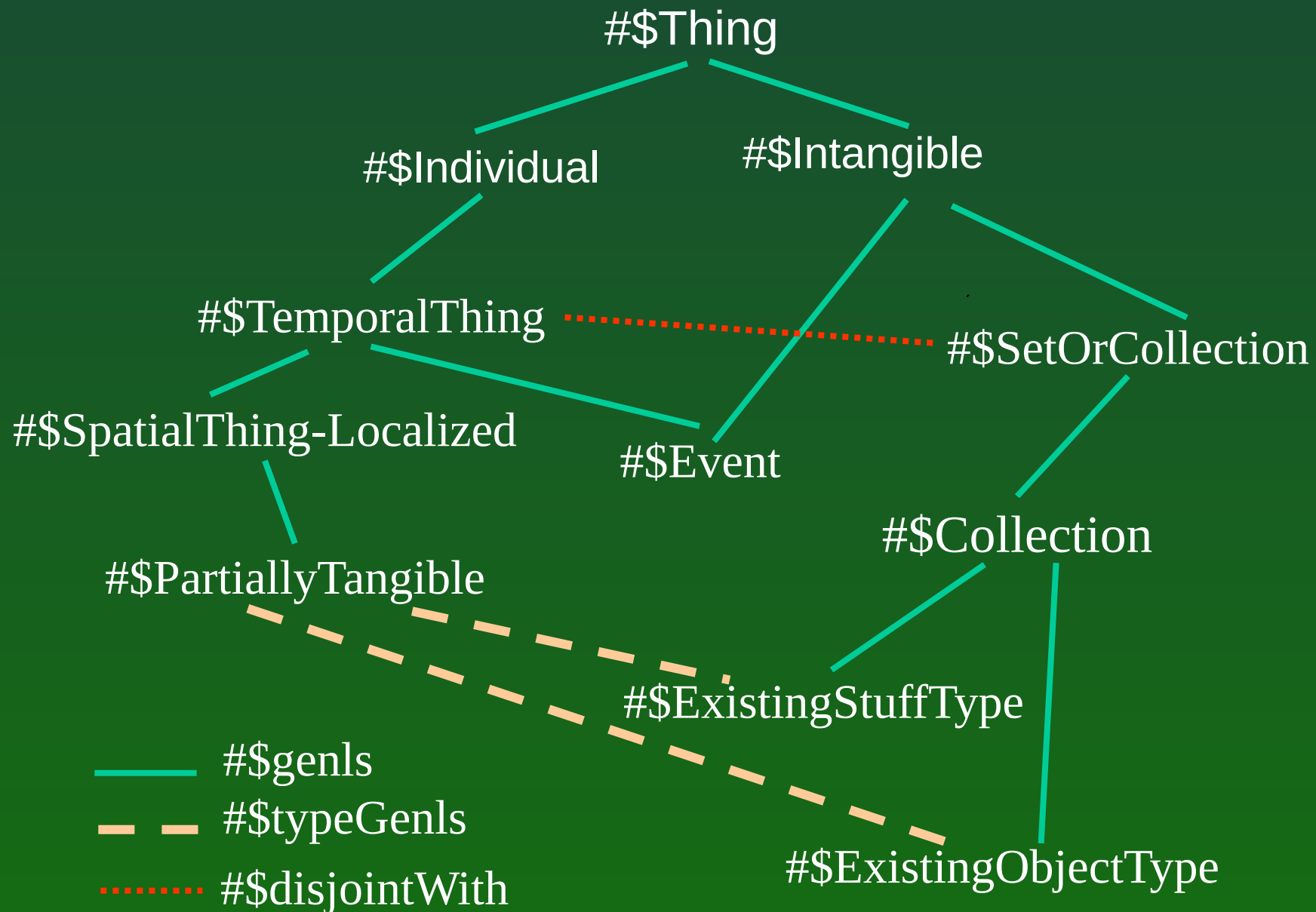
Asercje

- Wiedza deklaratywna, wnioskowanie częściowo proceduralne.
- CycL wykorzystuje logiczne stwierdzenia (*assertions*).
- CYC ma >10.000 predykatów, ok. 25 mln asercji, prostych faktów, reguł i sposobów wnioskowania; stwierdzenie może wiązać się z 10-100 innymi, stąd liczba powiązań w tym systemie sięga milionów.
- Pierwsze > 1.5M reguł/powiązań odnosi się do globalnej ontologii.
Jak reprezentować podstawowe pojęcia w bazie wiedzy?
- Opis + relacje najczęstszych sposobów użycia danego pojęcia.
- Zamiast zredukowanego opisu próbuje się opisać wszystkie sytuacje.
- CYC to pierwszy program wykorzystujący globalną ontologię, czyli klasyfikację bytów.
- Kwestie ontologii dyskutowane były od zarania filozofii, teraz to kwestia eksperymentalna.

Ontologia podstawowa

- Wszystko jest rodzajem rzeczy, konkretnej lub abstrakcyjnej. Rzeczy indywidualne i kolekcje, zbiory rzeczy.
- Indywidualne, np. *Jan, Polska, nos Jana*, mogą mieć części.
- Kolekcje, np. *osoba, naród, nos*, mogą mieć podzbiory.
- Rzeczy *nienamacalne* nie mają masy: *zdarzenia, liczby, prawa*.
- Rzeczy *namacalne* mają masę: *ciało człowieka, jabłko czy kurz*.
- Rzeczy złożone - cechy namacalne i nienamacalne,
np. *osoba* ma *ciało* i *umysł*.
- Substancja to *ObiektIndywidualny*, pocięta zachowuje własności;
- Własności zewnętrzne (*extrinsic*)
wewnętrzne (*intrinsic*), zarówno części jak i całego obiektu.

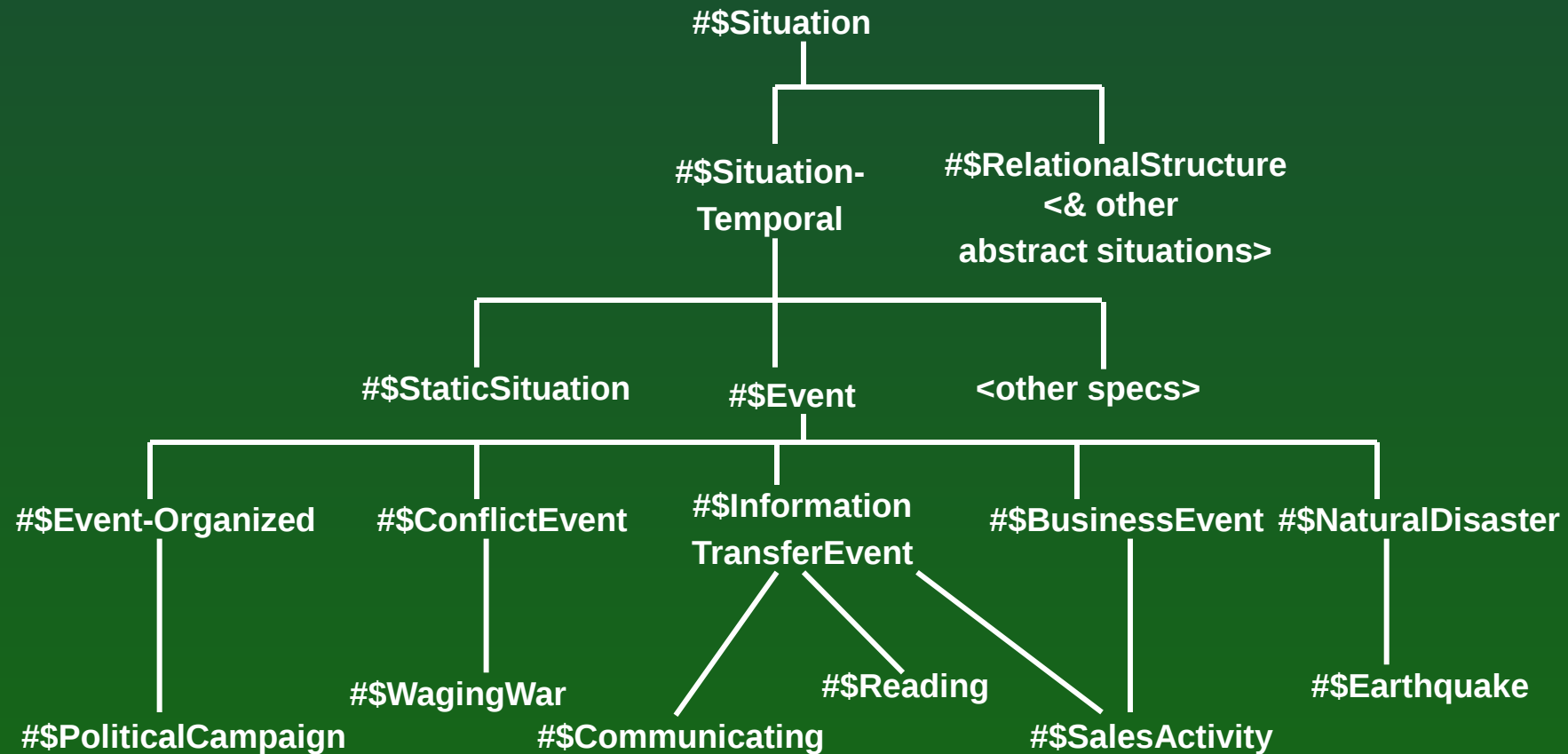
Kolekcje na najwyższych poziomach



Ontologia cd.

- Zdarzenia: rzeczy dziejące się w czasie.
- Procesy: zdarzenia, które po podzieleniu zachowują własności.
- *Spacer* jest procesem, ale *kilometrowy-spacer* nie jest.
- Zdarzenia: wł. temporalne, interwały czasowe i zbiory interwałów.
- *Slot*, czyli *szufladka*, to podklasa nienamacalnego.
- Szufladki definiujące, buchalteryjne (zapisujące informacje o pochodzeniu), ilościowe itp.
- **Agent**: obiekt złożony, zbiór inteligentnych istot, ludzie, firmy, programy.
- Agent ma przekonania (*beliefs*), może też przypisywać przekonania innym.
- Przekonania nie zawsze są prawdziwe, trzeba odróżnić lokalne przekonania agentów od własnej wiedzy o świecie.

Częściowa hierarchia zdarzeń



Cyc Ontology & Knowledge Base

Cyc contains (at 2009):

~13,500 Predicates

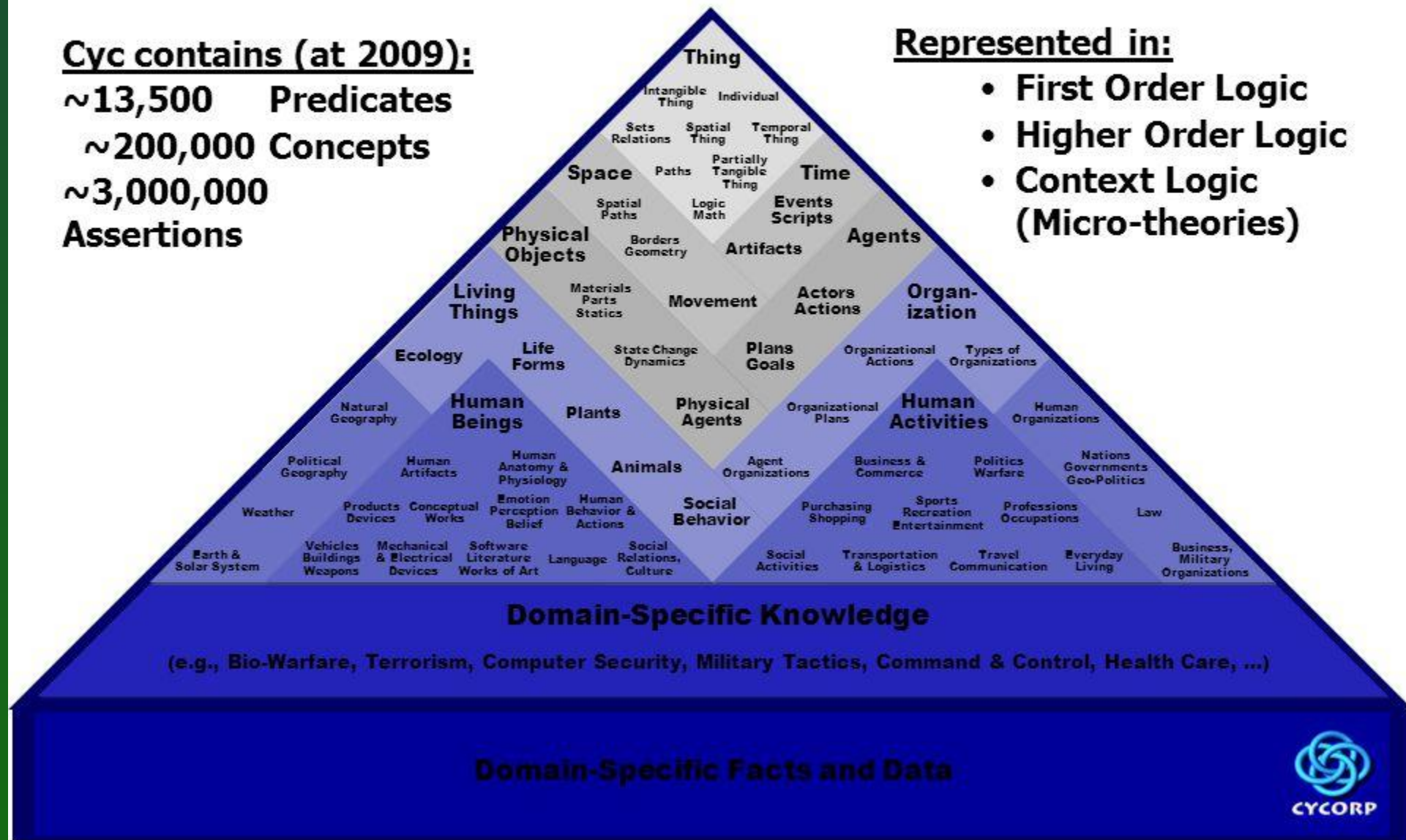
~200,000 Concepts

~3,000,000

Assertions

Represented in:

- First Order Logic
- Higher Order Logic
- Context Logic (Micro-theories)



Ontologia cd.

- „Upper Cyc ontology” to 6000 najbardziej ogólnych pojęć podstawowych „human concensus reality”, dostępne w Internecie.
- Szczegółowe struktury dodawane są do bazy wiedzy Cyc od 1984 roku. Dodatkowa wiedza z dziedziny, do której stosuje się system Cyc dodawana jest przez użytkownika.
- Podstawowa ontologia jest: uniwersalna, niezbędna do konstruowania innych pojęć, zawiera rozróżnienia konieczne i wystarczające dla większości zastosowań.
- „Konieczne” - warto je zrobić z teoretycznego i pragmatycznego punktu widzenia.
- „Wystarczające” - wystarczająco dużo rozróżnień by umożliwić zastosowania: dzielenie się wiedzą, rozstrzygnięcie niejednoznaczności języka naturalnego, szukanie wiedzy w bazach danych.
- Budowa ontologii to duża gałąź AI, strona pokazująca ontologię CyC: The [UpperOntology of the Cyc Project](#) używa [hiperbolicznego drzewa](#).

Terminologia Cyc

- Term - stała, np., #\$Skin, nazwa pojęcia w KB.
- Stała = zbiór (np. zbiór wszystkich ludzi), obiekt indywidualny (konkretną osobę), słowo, kwantyfikator (np. 'istnieje'), relację (predykat, funkcję, szufladkę, atrybut), itp.
- Podstawowa ontologia - szczyt wielkiej góry pojęć KB.
- Opis stałej: nazwa pojęcia + komentarz pomagający zrozumieć znaczenie i sposób użycia tej stałej.
- Informacja dotycząca hierarchii, *isa*, zbiory i nadzbiory.
- Hiperpołączenia do innych pojęć, np. #\$TactileSensor.
- Pojęcie „skóra” reprezentuje jeden z obiektów, które zdefiniowane są jako zbiór kawałków, czyli jako kolekcję.

Terminologia cd.

- Fizjologia: definicja stałej `#$Skin`, czyli „Skóra”.
- (Kawałek) skóry służy jako zewnętrzna ochrona i czujniki dotykowe pokrywające (część) ciała zwierzęcia. Skóra jest zbiorem wszystkich kawałków skóry. Niektóre przykłady to:
- `#$TheGoldenFleece` („Złote Runo”, reprezentujące całą skórę zwierzęcia) i (`#$BodyPartFn` `#$YulBrynnner` `#$Scalp`) (reprezentujący małą część jego skóry).
- isa: `#$AnimalBodyPartType`
- genls: `#$BiologicalLivingObject` `#$AnimalBodyPart` `#$SheetOfSomeStuff`
- `#$VibrationThroughAMediumSensor` `#$TactileSensor`
- `#$Skin` należy („isa”, czyli jest elementem) do zbioru „części zwierząt”, przykłady kojarzące się ze skórą - skóra na łysinie znanego aktora, oraz nadzbiory (genls, czyli jest podzbiorem) zawierające skórę. X = skóra może być „płatem czegoś” lub „czujnikiem wrażliwym na dotyk”.

Bardzo skomplikowany opis, ale nie wystarcza by we wszystkich możliwych kontekstach prawidłowo zinterpretować pojęcie skóry, dopiero LLMy potrafią to zrobić stosując reprezentacje wektorowe.

Terminologia cd.

- #Collection - zbiory, klasy obiektów o wspólnych własnościach. Definicje kolekcji są hierarchiczne, tylko specyficzne generalizacje.
- Zmniejsza to bazę faktów, wydłuża rozumowanie.
- #Relation, zbiór wszystkich relacji obejmujących funkcje i predykaty. Należy określić typ argumentów, np. *isa* odnosi się do kolekcji.
- #mother : <Animal> <FemaleAnimal>
- (#mother ANIM FEM) oznacza że #FemaleAnimal FEM jest rodzaju żeńskiego
- parent of the #Animal ANIM.
- isa: #FamilyRelationSlot #BinaryPredicate
- Komentarz wyjaśnia, że definicja dotyczy tylko biologicznego sensu słowa matka, a nie metaforycznego, dotyczy tylko zwierząt.
- CycL: wiele sposobów ograniczania rodzaju argumentów predykatów.

Terminologia cd.

- Wykorzystanie relacji: matką osoby X jest osoba Y
- Wyprowadzenie ogólnych reguł, np.

Jeśli matką X jest Y to Y urodziła się przed X i Y była prawdopodobnie żywa w momencie narodzin X .

- Funkcje w Cyc są podobne do predykatów, zwracają dowolne stałe.

Np. nazwa osobnika rodzaju żeńskiego danego gatunku:

- `#$FemaleFn ($Jeleń)` = łania. Pozwala to uniknąć definicji pojęcia „łania”.
- Wartości i argumenty funkcji ograniczone są przez predykaty:
- `#$resultIsa`, „rezultat należy do”
- `#$resultGenl`, „rezultat jest podzbiorem”.
- Wynik działania `#$FemaleFn` musi należeć do zbioru `#$FemaleAnimal`, musi też być podzbiorem zwierząt.

Przykład

- Animals sleep at home.
- `(ForAll ?x (ForAll ?S (ForAll ?PLACE
(implies (and
(isa ?x Animal)
(isa ?S SleepingEvent)
(performer ?S ?x)
(location ?S ?PLACE))
(home ?x ?PLACE))))))`
- Jeśli x jest zwierzęciem i
jeśli jest wykonawcą „sleeping event”, czyli śpi,
to miejscem tego spania jest dom x.
- Konstrukcje te nie są naturalne, ale w planach jest tłumaczenie z angielskiego na CyCo-logiczny.

CYC - rozwój

- CYC jak umysł, ciągle się rozwija.
- W sumie > 2000 osobo-lat pracy, kosztował ok. 60 M\$, już w 2002 roku zawierał 300.000 pojęć + 3 miliony reguł; >25 mln reguł w 2020 roku.
- Początkowo napisany był w Lispie na stacjach Symbolics, teraz jest w C pod Unix, MacOS i Windows.
- Skalowalność systemu, przetwarzanie rozproszone.
- Stworzenie ontologii - wysiłek wielu osobo-wieków.
- Nowe ontologie, np. militarne, polityczne, ekonomiczne, naukowe.
- Testowanie wnioskowania na dziesiątkach milionów przykładów.
- Kategorie i funkcje, które wydawały się przydatne potem odrzucono, np. `#$CompletelyTangibleObject` czy `#$TransfersThrough`); niektóre kategorie niezbyt przyjemnie wyglądające, np. `#$SiblingDisjointCollection` czy `#$SkolemFunction` okazały się bardzo przydatne i przetrwały wszystkie testy użyteczności.

OpenCyc

Otwarta, publicznie dostępna wersja systemu (od 2002 roku) o nazwie [OpenCyc](#), składająca się w wersji 4.0 z:

- 239.000+ pojęć i ponad 2 milionów faktów dotyczących tych pojęć, częściowo definiujących, ograniczających i definiujących wzajemne relacje, których jest 26.000+.
- Modułu wnioskującego Cyc Inference Engine.
- Przeglądarki bazy wiedzy Cyc Knowledge Base Browser.
- Narzędzi do akwizycji wiedzy ze specjalistycznych dziedzin przez dialog z ekspertem w języku naturalnym.
- Dokumentacji systemu, specyfikacji języka CycL, specyfikacji interfejsu użytkownika Cyc API i innych.
- Kilku programów demonstracyjnych.
- Prawie cały CyC był publicznie dostępny, fragmenty pozostały komercyjne.

4/2006, powstała CyC Foundation i publiczny serwer, ale od 2017 roku jest to system komercyjny.

OpenCyc można ściągnąć ze strony [projektu Clarin](#) EU.

Research CyC

Od 2006 roku była dostępna pełna wersja CyC dla celów badawczych.

System programów Research Cyc Release 1.1 (2017) zawierał:

- Ponad 630,000 pojęć w ontologii ogólnej.
- Ponad 7 mln asercji (faktów i reguł), ponad 38,000 relacji częściowo definiujących (wiążących, ograniczających) koncepcje.
- Skompilowana wersja Inference Engine, Knowledge Base Browser.
- Funkcje NLP i CycL-to-English.
- Narzędzia do formułowania zapytań tłumaczące na CycL.
- Szablony do wprowadzania wiedzy.
- Ontology Exporter do tworzenia plików ontologii w formacie OWL.
- Dokumentacja i tutoriale.
- Specyfikacja języka CycL, CycL-to-Lisp, CycL-to-C, itd.
- Specyfikacja Cyc API do tworzenia aplikacji ResearchCyc.

Tak duży system wymagał drogiego sprzętu, jego zastosowania były ograniczone.

Cyc Platform

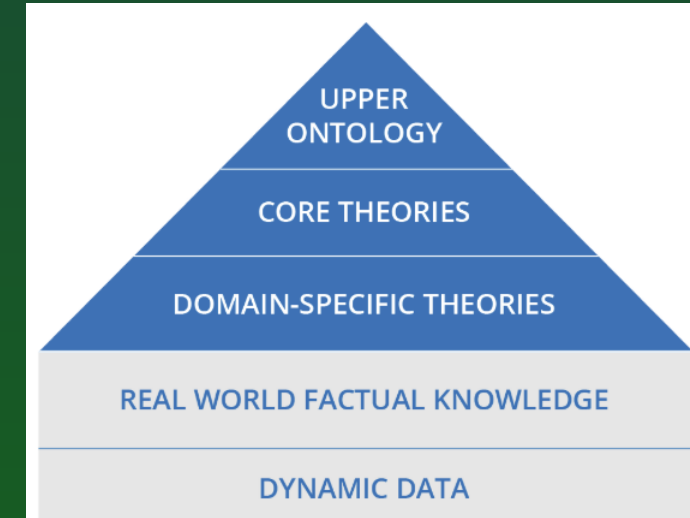
Cyc Foundation udostępniał wersję do badań o nazwie ResearchCyc, teraz jest [platforma komercyjna](#):

- Opis ponad 1.5 mln pojęć, 40 000 predykatów, 25 milionów faktów dotyczących tych pojęć, częściowo definiujących, ograniczających i definiujących wzajemne relacje, których jest >26.000.
- Ponad 1100 algorytmów modułu wnioskującego Cyc Inference Engine.
- Przeglądarki bazy wiedzy Cyc Knowledge Base Browser.
- Narzędzi do akwizycji wiedzy ze specjalistycznych dziedzin przez dialog z ekspertem w języku naturalnym.
- Dokumentacji systemu, specyfikacji języka CycL, specyfikacji interfejsu użytkownika Cyc API.
- Kilku programów demonstracyjnych.
- Dostęp wymaga licencji, która jest darmowa, ale od 2019 roku firma przestała udostępniać tą wersję, obiecując, że przygotują nową.
- Fundacja Cyc zapowiadała też Cyclopedię pozwalającą na przeszukiwanie Wikipedii używając pojęć ontologii CyC, ale projekt ten nie został zrealizowany.

CYC - zastosowania.



- Obecna lista zastosowań – medyczne, finansowe.
- Analiza, rozumienie i generacja tekstów w językach naturalnych.
- Semantyczna integracja baz danych, tworzenie tezaurusów tech/med.
- Semantyczne szukanie informacji (WWW!), szukanie wiedzy.
- Sprawdzanie spójności wiedzy, integracja heterogenicznych baz danych.
- Budowa i wykorzystywanie modeli użytkownika różnych systemów.
- Modelowanie użytkowników programów i urządzeń technicznych – HCI, ergonomia.
- Inteligentne interfejsy programowe, reagujące na intencje użytkownika.
- Symulacja inteligentnych zachowań postaci w grach komputerowych.
- Inteligentna symulacja wirtualnej rzeczywistości.
- Zarządzanie szpitalami.



Cyc-NL

- NLP wymaga wiedzy!

Jan widział jak samolot leciał nad miastem.

Jan widział góry gdy leciał nad miastem.

- Skąd wiemy, że „leciał” odnosi się raz do samolotu, raz do Jana?
- Cyc wie, że samoloty latają a góry nie - bez takiej wiedzy rozstrzygnięcie co czy kto leciał nie jest możliwe.

Cyc-NL zawiera:

- leksykon,
- parser syntaktyczny,
- interpreter semantyczny.

Cyc-NL

- Umożliwia analizę zdań złożonych i niejednoznacznych, zdań modalnych, negacji i zdań kwantyfikujących na kilku hierarchicznych poziomach, najczęściej jest to meta-rozumowanie heurystyczne.
- Komunikacja z Cyc za pomocą Cyc-NL pozwoli na używanie Cyc bez znajomości języka CycL.
- Cyc-NL: interfejsy baz danych w języku naturalnym, wykorzystania w tłumaczeniu maszynowym i analizie mowy ciągłej.
- CyCorp twierdzi, że ekspersyjność Cyc-NL jest taka jak języka angielskiego.
- Gra FACTory pomagała CyC określać prawdziwość faktów, zbierając odpowiedzi ludzi określała stopień prawdziwości różnych faktów.
- GECKA: Game Engine for Commonsense Knowledge Acquisition, Games with a purpose.

Heterogeniczne bazy danych

- W bazach danych jest informacja szczegółowa, kilka reguł może zawierać wiedzę pozwalającą streścić całą informację.
- Proces ten może wymagać rozumienia nazw atrybutów.
Np. atrybut „miasto” odnosi się do pojęcia #City.
- Opis struktury bazy danych w terminach CycL pozwala na zadawanie złożonych pytań bazie danych.
- Np. info. o ludziach, ich zawodach i miejscu zatrudnienia w jednej bazie + informacje o firmach i ich adresach, to można zadać pytanie:

Wypisz osoby z wyższym wykształceniem mieszkające nad Wisłą.

- Cyc wie, że pracownicy uczelni, zakładów badawczych, firm prawniczych, klinik medycznych mają zwykle wyższe wykształcenie, że ludzie zwykle mieszkają w pobliżu miejsca pracy oraz że „nad Wisłą” oznacza w pobliżu rzeki może odpowiedzieć na takie pytanie. Korzystając z Cyc-NL można je zadać bezpośrednio w języku angielskim, bez pisania programu w SQL.

Bazy danych

- Wiedza Cyc pozwala na formułowanie meta-pytań w stosunku do baz danych, używając hierarchii pojęć zawartej w ontologii.
- Możliwa jest konstrukcja raportów na podstawie pytań niezależna od struktury czy opisu samej bazy danych, na podstawie pytań.
- Wiedza pozwala na powiązanie dostępnych nazw z poszukiwanymi.

Np. szukamy zdjęcia „kogoś w niebezpiecznej sytuacji”.

- Cyc wie, że „strzelający terrorysta” lub „żołnierz w akcji” pasuje.
- Cyc + Cyc-NL w systemie wyszukiwawczym dla setek tysięcy zdjęć. Opisy zdjęć => formuły w języku CycL.
- Cyc-NL tłumaczy z ang. na CycL. Np. „przestraszona osoby”:

(#\$and (\$isa ?x #Person)

(\$feelsEmotion ?x #Fear #High))

CYC - zastosowania cd.

- Cycorp udostępnia Cyc w całości lub części, np. w systemach SAIC, Teknowledge i innych firm.
- Nowa koncepcja łączy się z innymi, podaje się stwierdzenia łączące ją z innymi koncepcjami - jeśli jest ich dostatecznie dużo wystarczy płytke wnioskowanie.
- Nowe koncepcje muszą być prawidłowo usytuowane w stosunku do podstawowych by dwa systemy można było zintegrować.
- Wiedza CYC ma strukturę sformalizowanych powiązań, mikroteorii.
Różne poziomy szczegółowości, różnice punktów widzenia, różnice kulturowe, czasu w którym powstała ocena itp.
- 10 różnych modeli ekonomicznych USA jako mikroteorie, których koncepcje odnoszą się do podstawowej ontologii.
Te same terminy wewnątrz mikroteorii mają różne definicje, podlegają innym regułom, ale mają zbliżone odniesienia do podstawowych koncepcji ontologicznych.

CycSecure

Program symulujący ataki na sieci komputerowe.

- Program składa się z:
- programu symulującego sieć komputerową na podstawie analizy parametrów rzeczywistej sieci.
- bazy wiedzy o zagadnieniach bezpieczeństwa sieci, ciągle uzupełnianej o nowe informacje
- analizatora planów ataku, odgrywającego rolę hackera dokonującego ataków na symulowaną sieć.

Zalety: nie zakłóca działania sieci;

- pozwala na testowanie planów, sekwencji działań;
- działa systematycznie, automatyzuje przyrost nowej wiedzy.

CYC - friends.

2 systemy CYC tworzą „The Cycic Friends Network”, protokół wymiany informacji w sieci (KQML).

- Wspólna wiedza podstawowa, The Core Cyc KB, 0.5 Ma.
- Wiedza specjalistyczna, np. geografia i polityka.
- Rozwiązują problemy przekraczające możliwości każdego z nich.

Przykładowy problem:

- „Które państwa Bliskiego Wschodu mogą zaatakować inne kraje ?”
- Odpowiedź: „Irak może zaatakować Kuwejt.”
- Fakty: Irak ma autorytarny rząd. Irak ma silną armię. Kuwejt ma słabą armię. Irak graniczy z Kuwejtem. Irak leży na Bliskim Wschodzie.
- Jeśli państwo ma rząd autorytarny i silną armię oraz graniczy z państwem, które jest militarne słabe, to jest duża szansa, że je zaatakuje.

CYC - friends 1

„Wymień jakichś demokratycznie wybranych przywódców krajów leżących na północ od równika”.

- Odpowiedź: „John Major” , ... bo mamy fakty:
- Jeśli kraj ma demokratyczny rząd to przywódca został demokratycznie wybrany. Wielka Brytania ma demokratyczny rząd.
- John Major jest premierem rządu W. Brytanii.
- Jeśli rejon A jest częścią rejonu B a rejon B jest na północ od C to rejon A jest na północ od C.
- Europa leży na półkuli północnej.
- Półkula północna leży na północ od równika.
- Wielka Brytania leży w Europie.
- Odpowiedzi wymagają wiadomości geograficznych + politycznych.
- Mamy tu rozumowanie ogólne (G.B. na północ od równika) i proceduralne.

CYC - friends 2

Agent Cyc mający trudności rozgląda się za pomocą - innym agentem.

- Jeśli nie potrafi go znaleźć to zwraca się do Cyc-brokera informacji. Proces konsultacji może wymagać kooperacji wielu agentów.
- Prototyp rozproszonej architektury i 3 agentów: GeoAgent, PolAgent i EcoAgent - znają się na geografii, polityce i ekonomii.
- Automatyczna rozbudowa bazy wiedzy Cyc w określonych domenach.
- **Baza informacji o filmach** (Cardiff Database) zawiera informacje o aktorach, reżyserach, producentach, scenariuszach. Struktura nie jest regularna, dokumenty w html.
- Agent-ekspert filmowy odwołuje się do tej bazy a inne systemy Cyc, potrzebujące takiej wiedzy, odwołują się do tego agenta.
„W jakich filmach grał Roland Reagan?”
- CycL `#$actedInMovie #$RonaldReagan ?x`
Dane giełdowe, profile firm, informacje z CIA World Fact Book.
- Niestety nie widać nowszych eksperymentów w tym kierunku.

Inne zastosowania



- Ponad 600 licencji; wiele zastosowań wojskowych (Lista zastosowań).
Wiele rozwiązań wymagających integracji (cross-vertical business challenges).

- Cross-Vertical Solutions

- Energy Pre-Spud Well Planning Virtual Drilling Advisor

- Virtual Completions Advisor

- Healthcare Patient Care and Throughput Monitor

- Financial Services Virtual Financial Analysis

- Compliance Monitoring and Reporting

- Conversational Training Tool

- Resource Planner Supply Chain Optimization

- Enterprise Knowledge Layer

- >200 szpitali

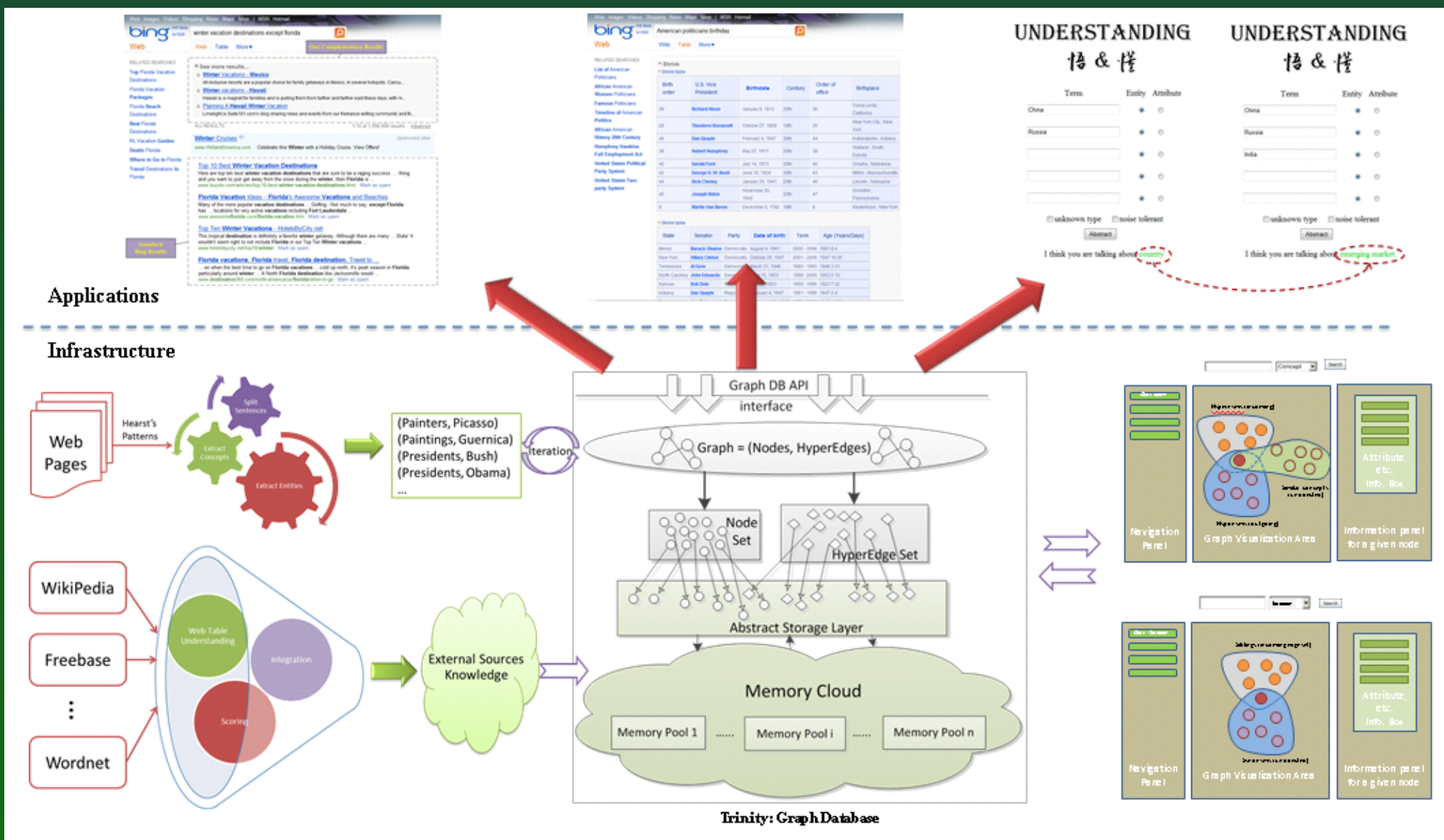


Duże ontologie open-domain

name	# of concepts	# of isA pairs
• ResearchCyc ≈	120,000	< 5,000,000
• Probase (MS)	2,653,872	20,757,545
• Freebase	1,450	24,483,434
• YAGO	352,297	8,277,227
• DBPedia	259	1,900,000
• WordNet	25,229	283,070
• WikiTaxonomy	111,654	105,418
• OMCS	173,398	1,030,619

To tylko nieliczne z ontologii, które powstały.

MS Probase + ConceptGraph





- HPKB, High Performance Knowledge Bases, inicjatywa DARPA.
- Projekt rozpoczęty w 1997 roku, ma znaczenie historyczne. Rapid Knowledge Formation (DARPA 1999-2013), Digital Aristotle (project Halo), collaborative Assitant for Rapid Knowledge Formation.
- Bazy wiedzy przydatne do problemów strategii militarnej i politycznych sytuacji kryzysowych. Problemy stanowiące wyzwanie dla systemów wspomagania decyzji.
- Wspomaganie podejmowania decyzji, analiz sytuacji międzynarodowej, przewidywanie kryzysów, ostrzeganie przed kryzysami, wspomaganie kształcenia politologów i wspomaganie modelowania historii kryzysów.
- Scenariusz kryzysu w zatoce Perskiej.
- Scenariusze kryzysów na Bliskim Wschodzie.
- Zarządzanie jednostkami wojskowymi na polu bitwy.

Niestety nadal takie systemy nie są stosowane, politycy podejmują złe decyzje samodzielnie.

IBM WATSON

Watson Q/A

- System QA firmy IBM. Projekt rozpoczęty w 2007 roku.
- W 2011 [IBM Watson](#) wygrywa z dwoma mistrzami teleturnieju Jeopardy (Va Banque). W konkursie ramię robota naciskało przycisk.
- W 2014 IBM utworzył Watson Group. Zaczęła się era “kognitywnych obliczeń”, od zabawy do zastosowań medycznych i AI dla biznesu.
- “Massively parallel probabilistic evidence-based architecture”.
- Początkowo odpowiedź wymagała 3 godzin, skrócono ją do 3 sek.
- Duże wymagania obliczeniowe.
- Wejście czysto tekstowe, brak analizy mowy czy obrazu.
- [Watson tutorials](#).

Watson

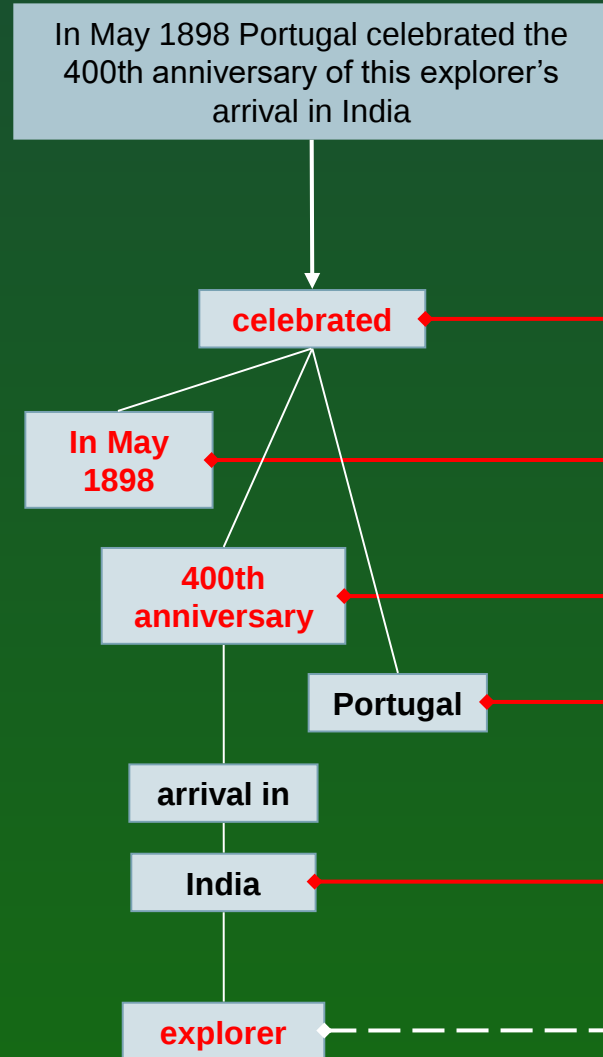
- Podstawą jest otwarty framework IBM UIMA: unstructured information management application; nadaj strukturę informacjom.
- Wszystkie zasoby tekstowe zostały zindeksowane wcześniej, nie używano szukania w czasie gry.
- Użyto wielu różnych źródeł informacji.
 - Podstawą była Wikipedia, Project Gutenberg, tezaury, doniesienia gazetowe itp.
 - Dodatkowe informacje aktywnie szukane w sieci i dodane do korpusu treningowego.
 - Tekst bez anotacji i bez struktury.
 - Wykorzystano analizę struktury Wikipedii z projektu Yago, oraz analizę pojęć z Wordnet.
- Jedynie 2% odpowiedzi w Jeopardy! było bezpośrednio dostępne w utworzonym korpusie, pozostałe wymagały wnioskowania.
- Klucz: integracja różnych źródeł informacji i technik wnioskowania.

Składowe DeepQA

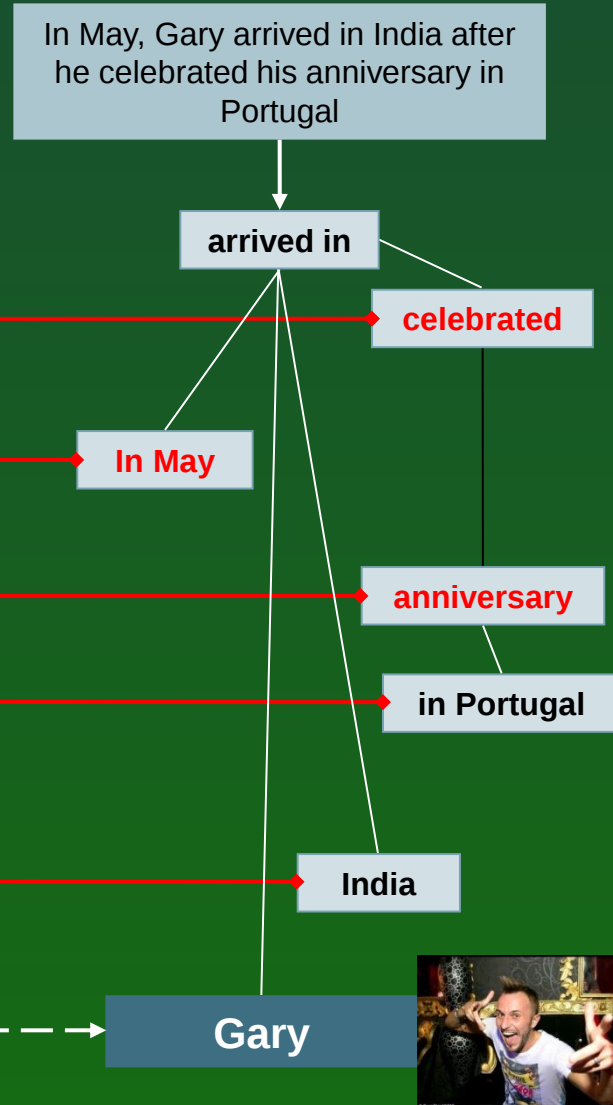
- Około 100 różnych technik.
- Akwizycja: utworzenie korpusu (offline), zabawa w gry słowne.
- Preprocessing
- Liczne narzędzia NLP, parsery
- Wyszukiwanie odpowiedzi, mechanizmy wnioskujące
- Oceny jakości odpowiedzi.
- Strategie gry
- Strategia wciskania przycisku po osiągnięciu pewności odpowiedzi.

Złożone pytania

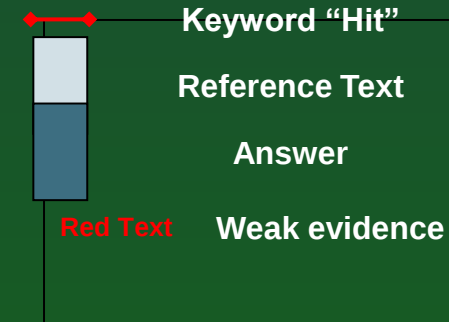
Question:



Supporting Evidence:



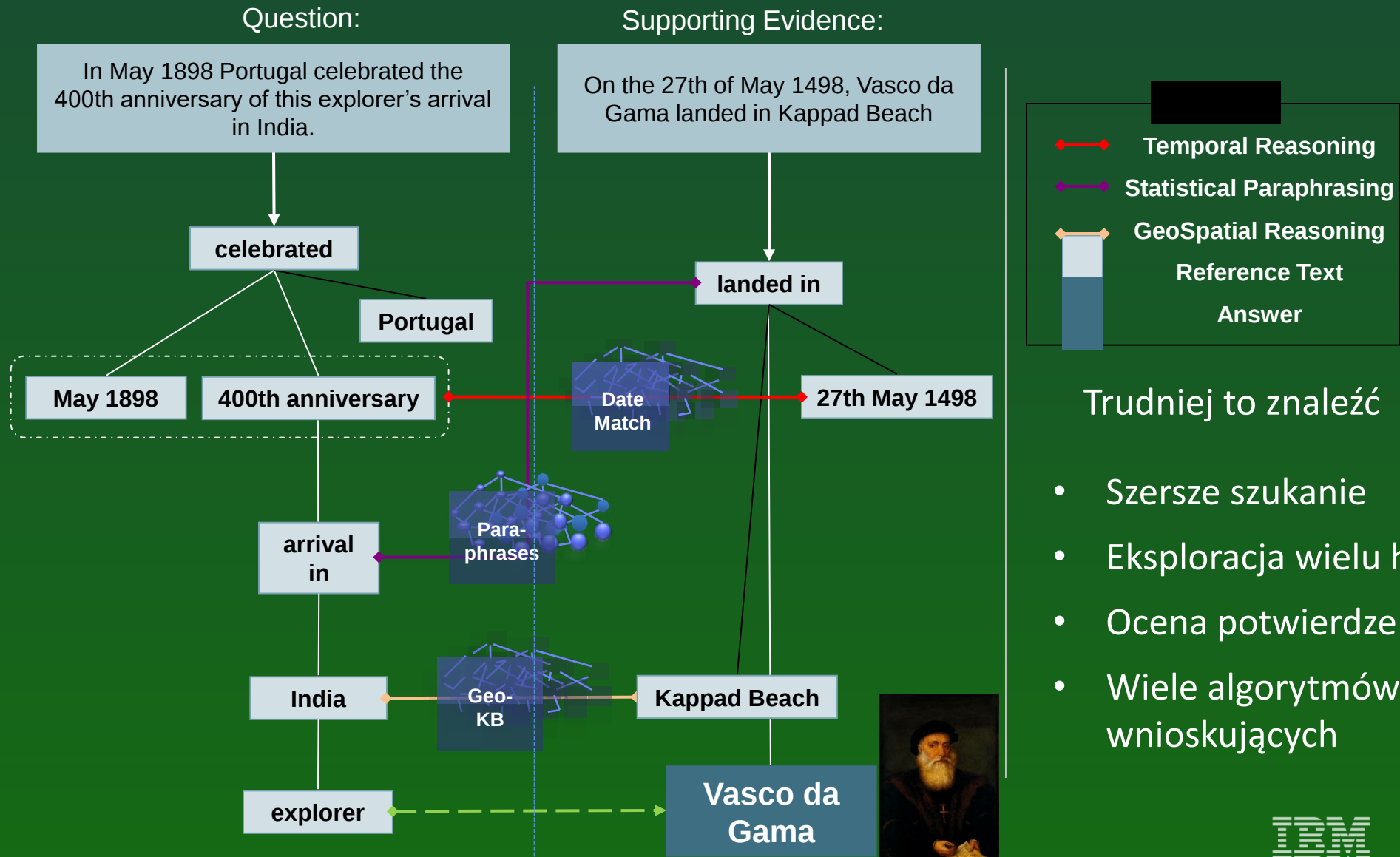
Legend



Hipoteza: odpowiedzią jest "Gary" ale system nie może się opierać tylko na dopasowaniu słów kluczowych bo to słabe potwierdzenie.

IBM

Watson używa głębszej analizy



Preprocessing

- Określenie kategorii pytań
 - faktoidy
 - rozkładalne na dwa lub więcej członów
 - puzzle
- Określenie kategorii odpowiedzi (lexical answer type, LAT)
 - film? osoba? miejsce? piosenka? nowela?

Około 2500 z 20,000 pytań należało do tych kategorii.

Analiza NLP Watsona

- Parsery gramatyczne
- Dekompozycja na prostsze fragmenty
- Tagi semantyczne składowych pytań
- Rozstrzygnięcie ko-referencji
- Odkrywanie relacji między słowami
- Rozpoznanie nazw własnych (named entity)

Szukanie przydatnych tekstów

- Wspecjalizowane szukarki: Indri (projekt Lemur), Apache Lucene, SPARQL RDF Query
- Szukanie składowych pytań podobnych do użytych w wyszukiwaniu informacji z sieci
- Skupienie na odzyskaniu (recall) wszystkiego, niska czułość.
- Generowanie setek hipotez
- Dla dobrze określonych kategorii LATs (np. nazwy krajów, stanów USA, prezydentów) lista możliwych odpowiedzi generowana jest bezpośrednio.
- Dla pozostałych konieczne jest wnioskowanie

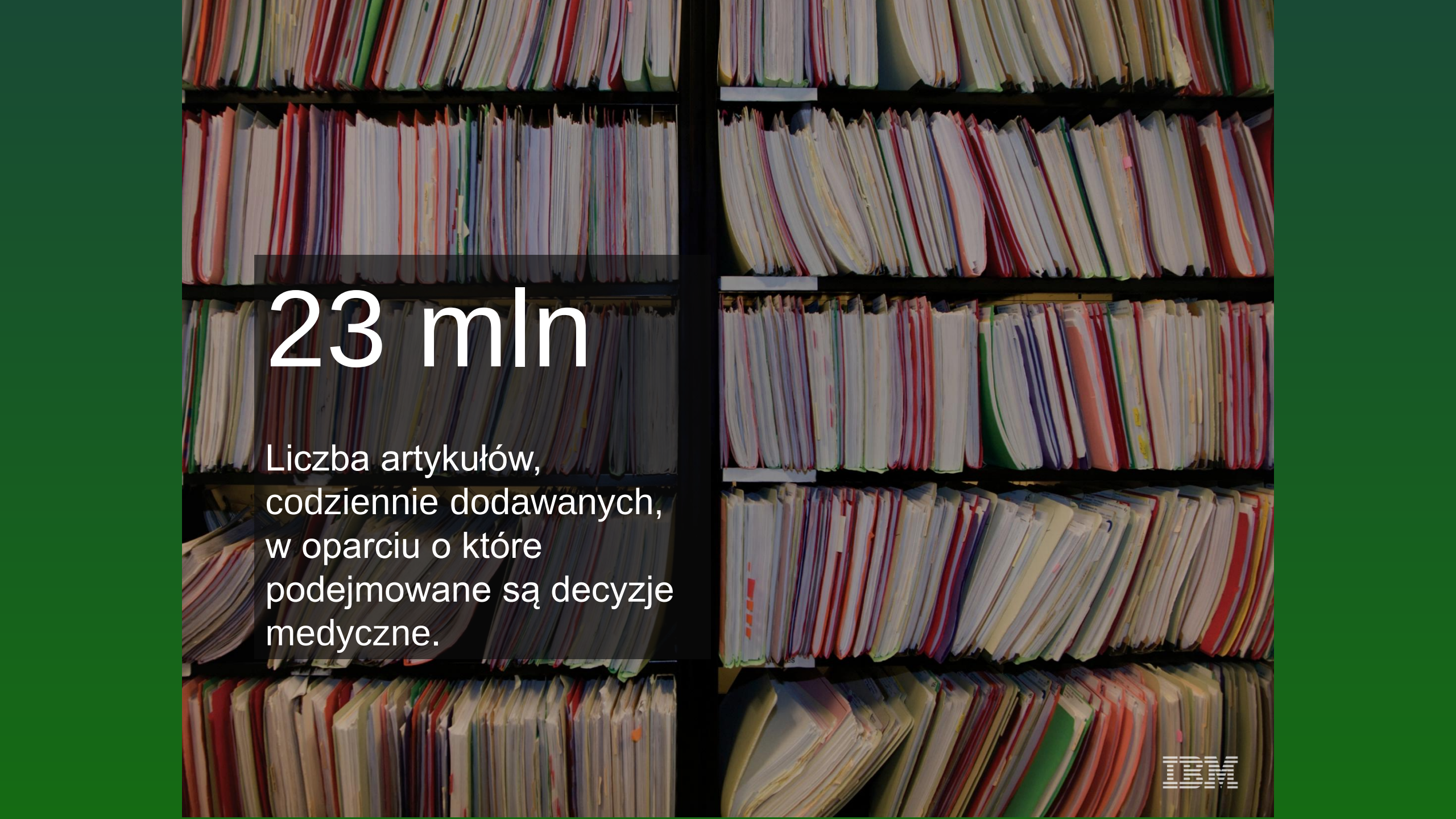
Ocena hipotez

- Ocena wygenerowanych hipotez przez
 - miękkie filtrowanie (soft filtering), szybkie algorytmy zostawiają około 100 hipotez
 - szukanie wsparcia dla każdej z hipotez, dodatkowe zapytania skierowane do korpusu
- Oceny odpowiedzi
 - ponad 50 różnych algorytmów
 - od prostego zliczania słów do złożonych form rozumowania uwzględniającego relacje czasowe i przestrzenne
 - tworzony jest profil potwierdzeń (evidence profile): zaufanie do źródeł, taksonomia, geoprzestrzenne i czasowe relacje ...
- Łączenie odpowiedzi
- Końcowy ranking i estymacja wiarygodności.

Strategie końcowe gry

- Kiedy się zgłaszać do odpowiedzi (naciśnięcie przycisku)?
 - Ocena pewności >50%
 - Ocena ryzyka uwzględniająca uzyskaną przewagę przy niższych wartościach.
 - Nie zgłasza się jeśli błąd może doprowadzić do przegranej.
- Ile i jak warto deklarować?

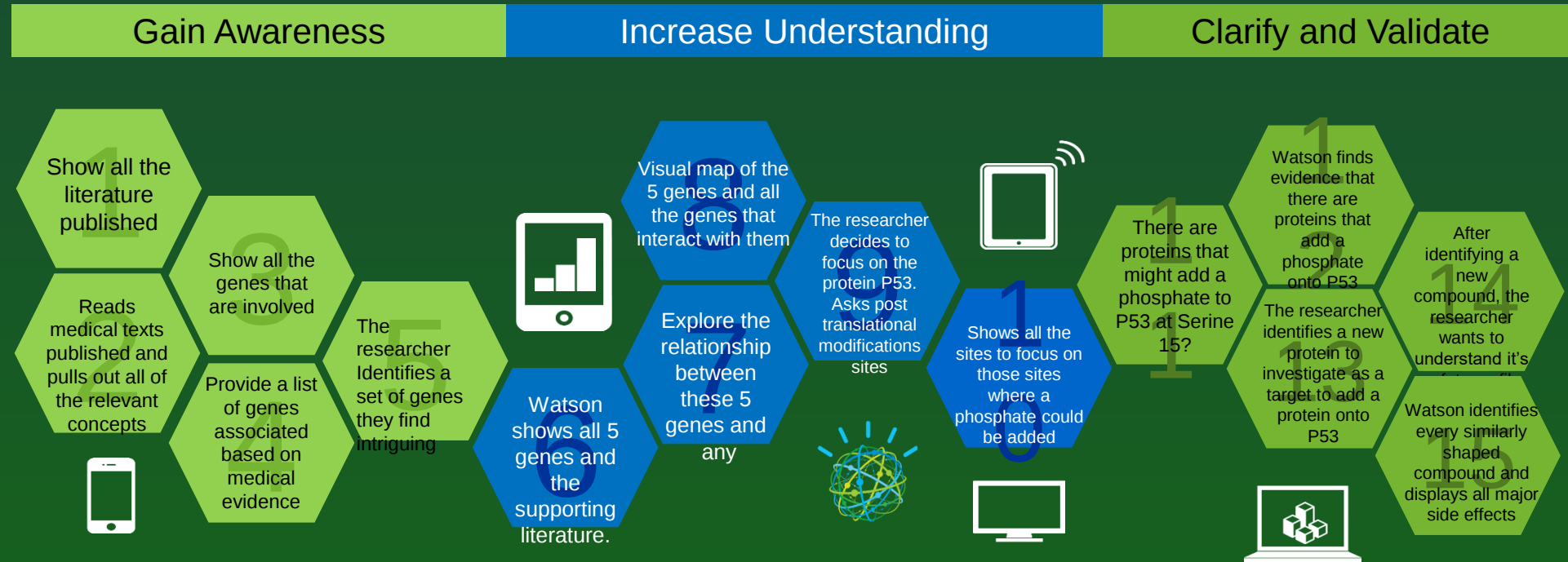




23 mln

Liczba artykułów,
codziennie dodawanych,
w oparciu o które
podejmowane są decyzje
medyczne.

Watson: Transforming the Discovery of Breakthrough Treatments



Watson Enhancements



Literatura

- Najwięcej informacji: <https://www.ibm.com/watson>
- Prezentacja [algorytmów IBM Watson](#) (ITT Bombay)
- Szczegółowy opis: AIMagazine-[Building Watson](#)-Overview of DeepQA Project
- Końcówka gry Jeopardy! https://www.youtube.com/watch?v=WFR3lOm_xhE
- [IBM Debater i Watson Discovery](#)
- Project Debater is the first AI system that can debate humans on complex topics, help people to build persuasive arguments and make well-informed decisions.
- Obecnie IBM połączył symboliczne możliwości Watsona z dużymi modelami językowymi, oferując system [WatsonX](#) dla dużych firm.

Przykładowe pytania

- Dlaczego Watson może się przydać onkologom? Dlaczego nie odniósł spodziewanych sukcesów?
- Po co stworzono CYC?
- Jaki zmierzyć jego wielkość?
- Jakie są podstawowe zasady konstrukcji CYC?
- Co wiesz o języku CyCL?
- Jakie mechanizmy wnioskowania w nim zaimplementowano?
- Co to jest ontologia? Globalna ontologia?
- Opisać podstawową ontologię CYC.
- Co daje się zrobić z CYC?
- Jakiego rodzaju zastosowania CYC umożliwia?