Course: Common Sense Reasoning

7. Manual Acquisition of Common Sense Knowledge

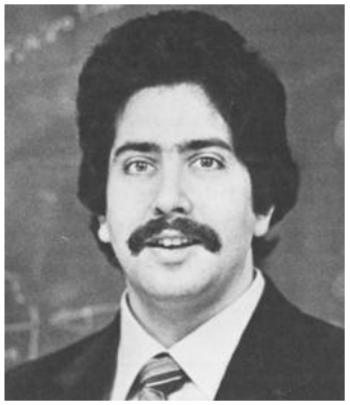
Martin Molina

Common sense knowledge can be acquired manually

- Knowledge engineers can build a common sense knowledge base in the following way:
  1. Analyzing manually different sources of knowledge related to common sense (e.g., texts, experts, etc.), and
  2. Writing the content of a knowledge base using an appropriate representation language (e.g., with a logic-based approach)
- Two representative cases are presented:
  - Cyc project
  - Sumo ontology

The Cyc project started in 1984

- The name Cyc is from encyclopedia
- Cyc started by Douglas Lenat
- Developed by the company Cycorp (since 1995)



Douglas Lenat

Cycorp®

[Lenat, 1995] [Lenat, Guha, 1990]

What is the goal of the Cyc project?

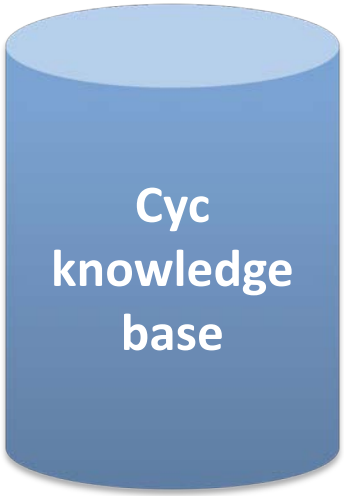
Codify millions of pieces of knowledge that compose human common sense:

People fly in planes

The water is wet

…

The project has generated a large knowledge base



7M assertions

0.5M terms

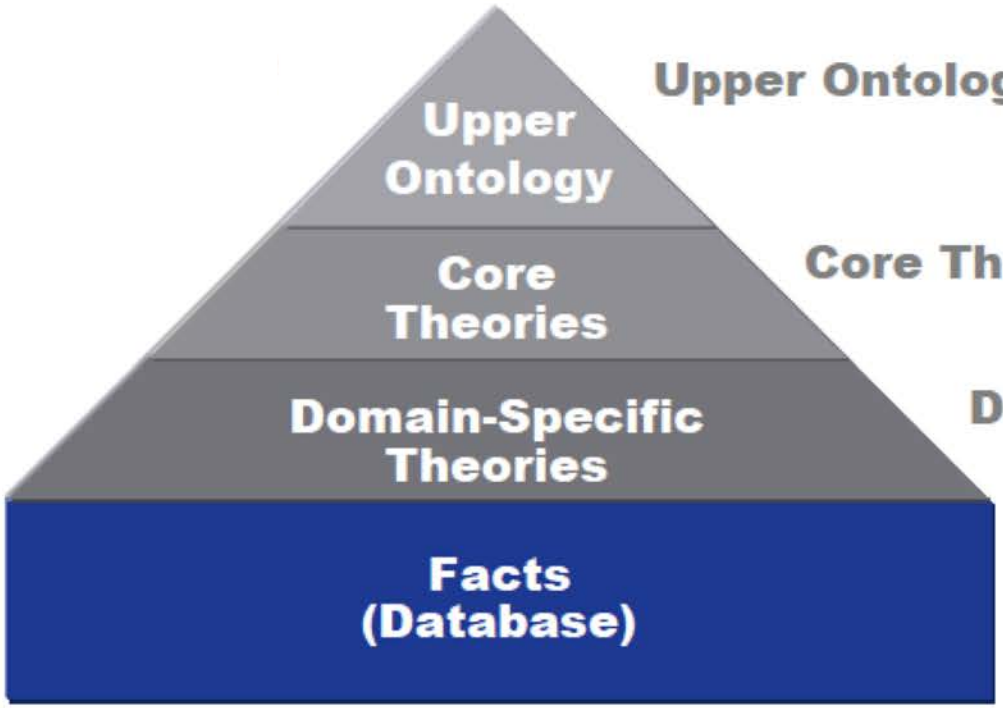
25K types of relations

Note: These numbers change periodically as result of the KB development and maintenance

What are words, terms and assertions?

- Words (thousands in usual vocabulary)
  - *Car*
  - *Person*
- Terms (hundreds of thousands)
  - *Fault tolerant computer*
  - *Developing country*
- Assertions (over a million)
  - *An electric guitar is a string musical instrument*
  - *People fly in planes*

The knowledge base is structured in layers



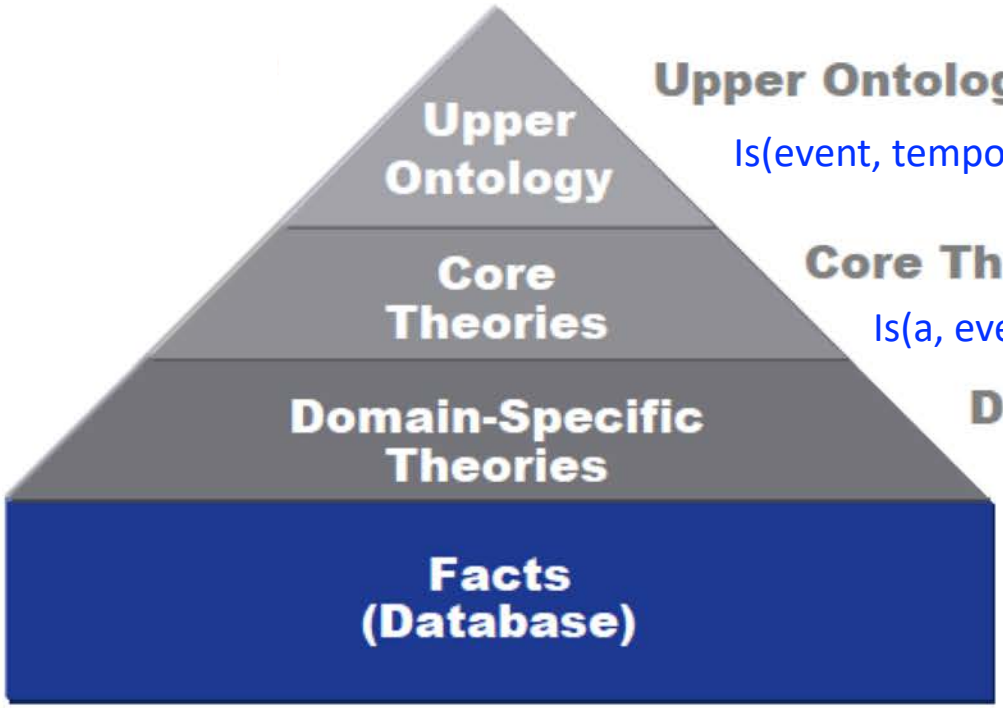
Upper Ontology: Abstract Concepts

Core Theories: Space, Time, Causality, ...

Domain-Specific Theories

Facts: Instances

The knowledge base is structured in layers



Upper Ontology: Abstract Concepts

Is(event, temporal-thing), Is(temporal-thing, individual)

Core Theories: Space, Time, Causality, …

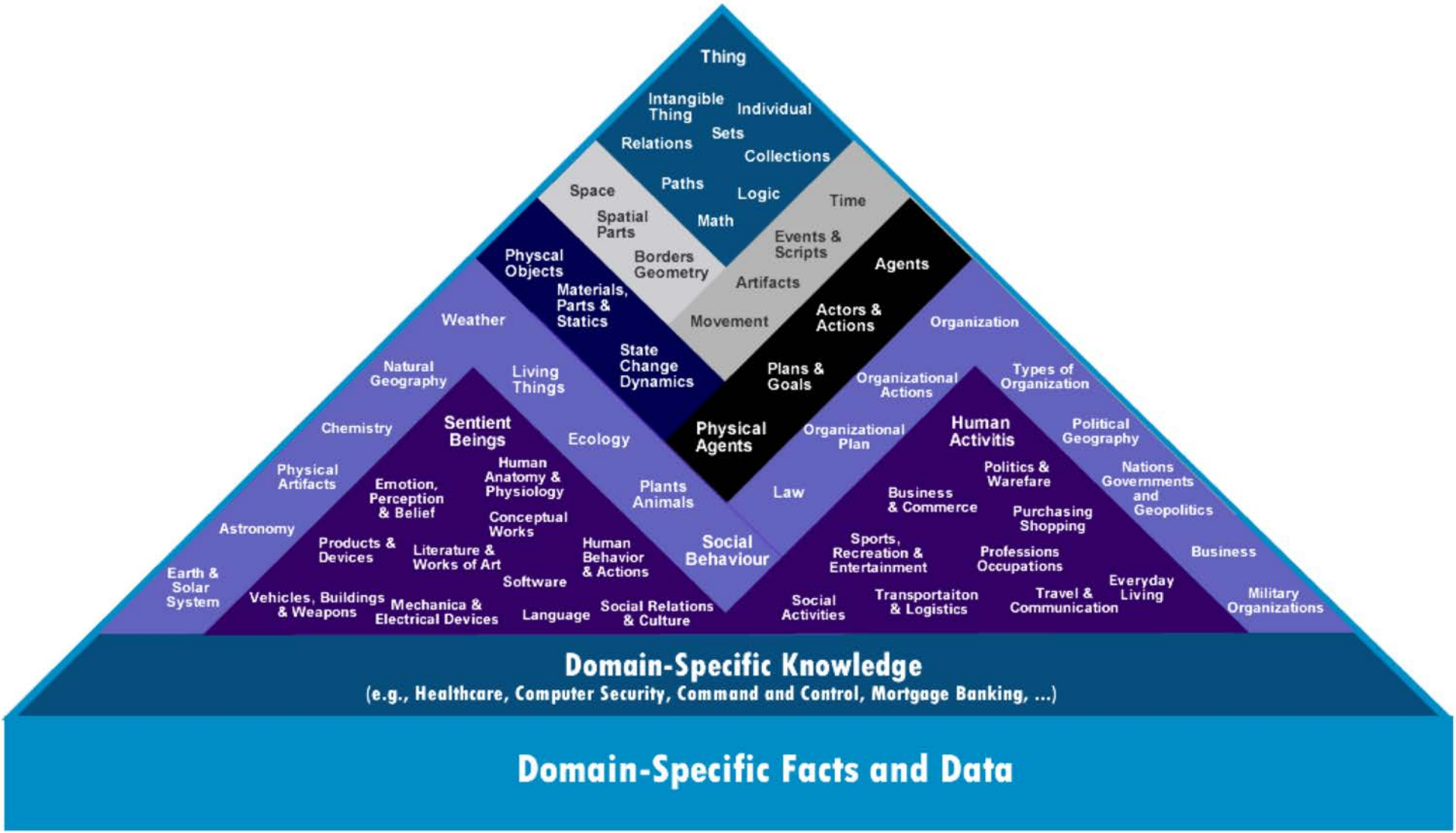
Is(a, event), Is(b, event), Causes(a, b) → Precedes(a, b)

Domain-Specific Theories

Is(x, bacteria), Is(y, animal) →
Causes(Exposed(y, x), Infected(y, x))

Facts: Instances

Is(John, person), Infected(John, anthrax)


Thing
Intangible Thing
Individual
Relations
Sets
Collections
Space
Paths
Logic
Time
Spatial Parts
Math
Events & Scripts
Physcal Objects
Borders Geometry
Agents
Materials, Parts & Statics
Artifacts
Weather
Movement
Actors & Actions
Organization
State Change Dynamics
Natural Geography
Living Things
Plans & Goals
Organizational Actions
Types of Organization
Chemistry
Sentient Beings
Ecology
Physical Agents
Organizational Plan
Human Activitis
Political Geography
Physical Artifacts
Human Anatomy & Physiology
Politics & Warefare
Nations Governments and Geopolitics
Emotion, Perception & Belief
Plants Animals
Law
Business & Commerce
Purchasing Shopping
Astronomy
Conceptual Works
Products & Devices
Literature & Works of Art
Human Behavior & Actions
Social Behaviour
Sports, Recreation & Entertainment
Professions Occupations
Business
Earth & Solar System
Software
Vehicles, Buildings & Weapons
Mechanica & Electrical Devices
Language
Social Relations & Culture
Social Activities
Transportaiton & Logistics
Travel & Communication
Everyday Living
Military Organizations
Domain-Specific Knowledge
(e.g., Healthcare, Computer Security, Command and Control, Mortgage Banking, ...)
Domain-Specific Facts and Data

How does Cyc represent knowledge?

- Cyc uses a logic-based representation
- Cyc uses a representation language called CycL

  [Matuszek, et al., 2006]

Constants are defined in CycL with the prefix #$

- Individuals
  - `#$BillClinton, #$Rover, #$DisneyLand` [Objects]
  - `#$likesAsFriend, #$bordersOn, #$objectHasColor` [Relations]
  - `#$and, #$not, #$implies, #$forAll` [Logical connectives, quantifiers
  - `#$RedColor, #$Soil-Sandy` [Attribute values]
- Collections
  - `#$Dog, #$SnowSkiing, #$PhysicalAttribute`

Sentences are used to form assertions and queries

```
(#$isa #$John #$Person)

(#$likesAsFriend #John #$Peter)
```

- A TruthFunction:
- is a relation that can be used to form sentences
- begins with a lower-case letter

- Types of TruthFunctions:
- Predicates: #$likesAsFriend, #$objectHasColor, #$isa
- Logical Connectives: #$and, #$or, #$not, #$implies
- Quantifiers: #$forAll, #$thereExists

Instances are represented using `#$isa`

`(#$isa X Y)` means:

"*X* is an instance of collection *Y*"

Examples:

```
(#$isa #$EiffelTower #$Tower)
(#$isa #$Canada #$Country)
(#$isa #$John #$Person)
(#$isa #$UnitedStatesMarineCorps #$ModernMilitaryOrganization)
```

Collections are represented using #$genls

(#$genls X Y) means:

"Every instance of collection X is also an instance of collection Y"

Examples:

```
(#$genls #$Dog #$Mammal)
(#$genls #$Tower #$FixedStructure)
(#$genls #$ModernMilitaryOrganization #$Organization)
```

Sometimes expressed as:

"Y is a genls (generalization) of X"

"X is a spec (specialization) of Y"

Logical connectives

```
(#$and
  (#$performedBy #$GettysburgAddress #$Lincoln)
  (#$objectHasColor #$Rover #$TanColor))

(#$or
  (#$objectHasColor #$Rover #$TanColor)
  (#$objectHasColor #$Rover #$BlackColor))

(#$not
 (#$mainColorOfObject #$Rover #$RedColor))

(#$implies
  (#mainColorOfObject #$Rover #$TanColor)
  (#$not (#$mainColorOfObject #$Rover #$RedColor)))
```

Quantifiers

Universal Quantifier

```
(#$forAll ?THING
  (#$isa ?THING #$Thing))
```

Everything is a thing

Existential Quantifier:

```
(#$thereExists ?JOE
  (#$isa ?JOE #$Poodle))
```

Something is a poodle

Others:

```
(#$thereExistsExactly 12 ?ZOS
  (#$isa ?ZOS #$ZodiacSign))
```

There are exactly 12 zodiac signs

```
(#$thereExistsAtLeast 9 ?PLNT
  (#$isa ?PLNT #$Planet))
```

There are at least 9 planets

Examples

There is at least one planet orbiting the Sun

$$\exists x[IsA(x, planet) \wedge Orbits(x, Sun)]$$

```
(#$thereExists ?PLANET
  (#$and
    (#$isa ?PLANET #$Planet)
    (#$orbits ?PLANET #$Sun)))
```

Everybody loves somebody

$$\forall x\ [IsA(x, person) \rightarrow \exists y\ [IsA(y, person) \wedge Loves(x, y)]]$$

```
(#$forAll ?PERSON1
  (#$implies
    (#$isa ?PERSON1 #$Person)
    (#$thereExists ?PERSON2
      (#$and
        (#$isa ?PERSON2 #$Person)
        (#$loves ?PERSON1 ?PERSON2)))
```

The fruit of the apple tree is red

$$IsA(x, Fruit(appleTree)) \rightarrow ColorOfObject(x, redColor)$$

```
(#$implies
  (#$isa ?X (#$FruitFn #$AppleTree))
  (#$colorOfObject ?X #$RedColor))
```

There are terms for domain specific relations

Biological relatives

```
(#$biologicalRelatives #$JerryLeeLewis #$JimmySwaggart)
```

Geographical subregions

```
(#$geographicalSubregions #$UnitedStates #$Utah-State)
```

Orbits

```
(#$orbits #$MoonOfEarth #$PlanetEarth)
```

Specific predicates can be used to define syntax and semantics of other predicates

arity

```
(#$arity #$biologicalMother 2)
```

arg1Isa, arg2Isa, ...

```
(#$arg1Isa #$biologicalMother #$Animal)
(#$arg2Isa #$biologicalMother #$FemaleAnimal)
```

The inference engine is a goal-directed theorem prover

Optimized for a large knowledge base:

- Shallow (incomplete)
  - The inference engine stops after a prefixed limit:
    - After T seconds, N depth steps, M chain rules, or N answers
    - Uses a small percentage of the knowledge base and not many steps
- Heuristics
  - Cyc has a set of heuristics (thousands) to rank potential inferences in a preference order.
- Contexts
  - Cyc's inference manages consistency with local inferences in microtheories.

Cyc has been used in multiple applications

Medical Records Analysis

Semantic interface to medical records databases

Intelligence / Counter-terrorism Analysis

Knowledge base about terrorism with a natural-language question-answering system

Financial Analysis

Questions that will impact their investment and trading decisions

TextPrism

Delivers information based on an individual‘s specific interests (with interest rules)

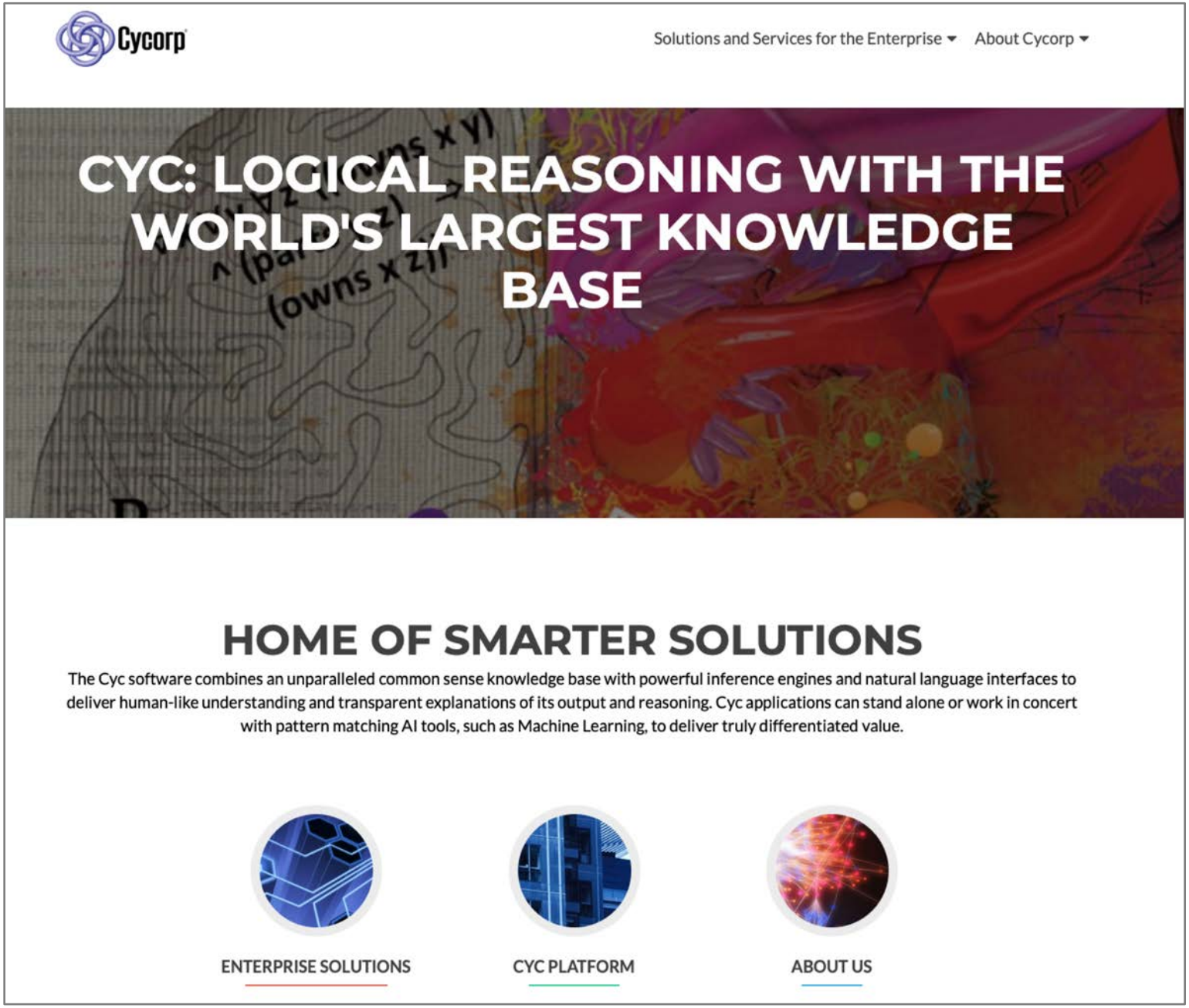
CycSecure

Security risk management and network intrusion prevention tool

OpenCyc is an open version of the Cyc knowledge base

- OpenCyc is part of Cyc
  - Only the core Cyc ontology
- Release 4.0 includes:
  - 2,100,000 triples and 240,000 terms
  - A java-based Cyc Inference Engine
  - The Cyc Knowledge Base Browser
- There is a more complete version for research purposes
  - ResearchCyc

https://www.cyc.com



http://dev.cyc.com

What are the strengths and weaknesses of Cyc?

- Strengths
  - The first attempt to build a common sense KB (1984)
  - Large content (millions of assertions about universals)
  - Expressive language and powerful inference (CycL, theorem prover)
  - Practical applications (terrorism, medicine, etc.)
- Weaknesses
  - Cyc is a commercial approach partially closed
  - It is incomplete
  - Unsatisfactory treatment of some concepts

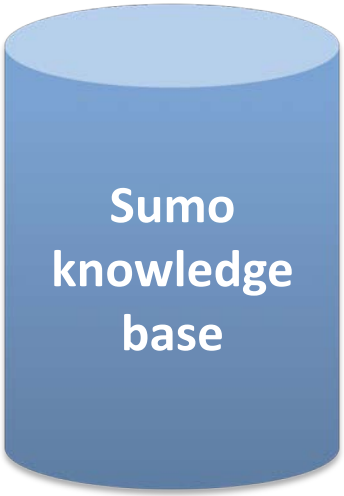
Sumo is a public knowledge base

- The name Sumo is from Suggested Upper Merged Ontology
- Developed initially in the company Teknowledge and then the company Articulate Softwre
- Adam Pease is the technical editor of Sumo

Adam Pease

[Pease et al., 2002]

The size of Sumo is small compared to the Cyc knowledge base



81,000 axioms

22,000 terms

5,000 rules

1,400 relations

Sumo uses the logic-based language SUO-KIF

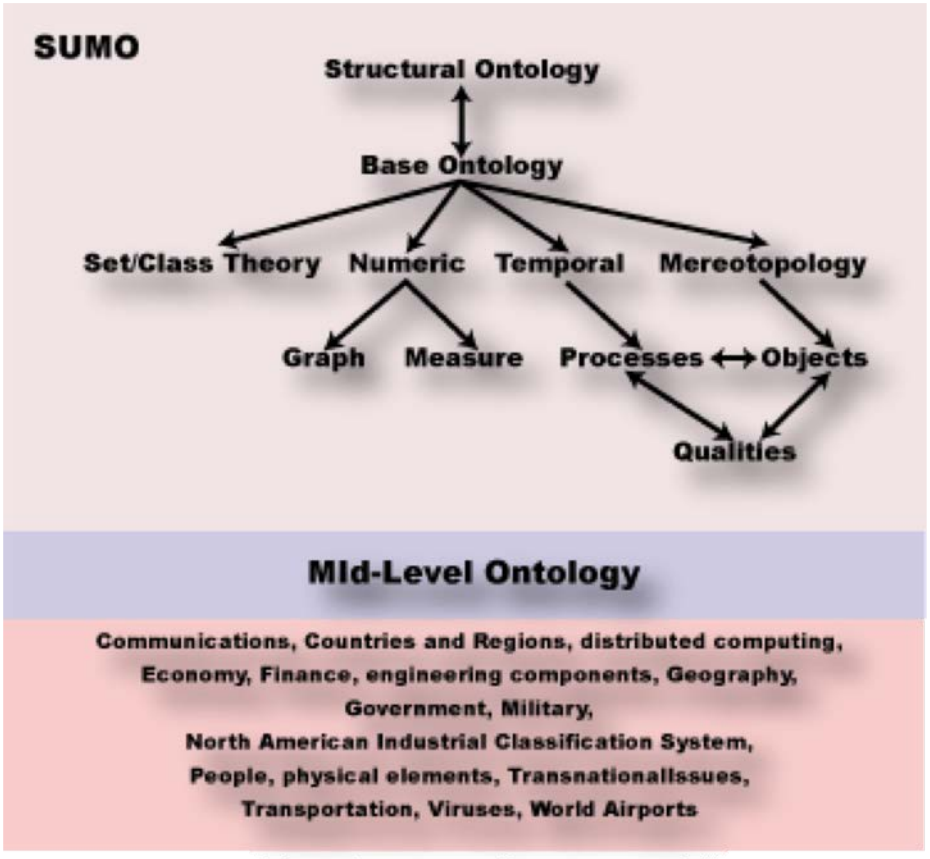
```
(forall (?X ?Y)
  (=>
    (and
      (instance ?X Farmer)
      (instance ?Y Tractor))
    (likes ?X ?Y)))
```

$\forall x, \mathrm{y}\ [Instance(x, farmer) \wedge Instance\ (y, tractor) \rightarrow Likes(x, y)]$

```
(=>
  (instance ?X Atom)
  (exists (?Y ?Z)
    (and
      (component ?Y ?X)
      (component ?Z ?X)
      (instance ?Y Proton)
      (instance ?Z Electron))))
```

$Instance(x, atom) \rightarrow \exists y\ \exists z\ [Comp(y, x) \wedge Comp(z, x) \wedge Instance\ (y, proton) \wedge Instance\ (z, electron)\ ]$

The knowledge base is structured in three main layers



Upper-level ontology

Mid-level ontology

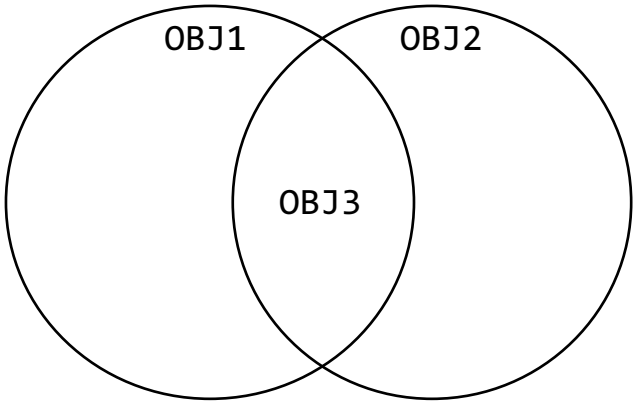
Domain ontologies

Extract from the upper-level ontology: relation “overlapsSpatially”

```
(subrelation overlapsSpatially connected)
(instance overlapsSpatially ReflexiveRelation)
(instance overlapsSpatially SymmetricRelation)

(<=>
  (overlapsSpatially ?OBJ1 ?OBJ2)
       (exists (?OBJ3)
            (and
                 (part ?OBJ3 ?OBJ1)
                 (part ?OBJ3 ?OBJ2))))
(=>
   (and
      (member ?OBJ1 ?COLL)
      (member ?OBJ2 ?COLL)
      (not (equal ?OBJ1 ?OBJ2)))
   (not
      (overlapsSpatially ?OBJ1 ?OBJ2)))

(=>
   (and
      (instance ?REL CaseRole)
      (instance ?OBJ Object)
      (?REL ?PROCESS ?OBJ))
   (exists (?TIME)
      (overlapsSpatially (WhereFn ?PROCESS ?TIME) ?OBJ)))
```



Extract from the mid-level ontology: subclass “Divorcing”

```
(subclass Divorcing Declaring)

(=>
  (instance ?D Divorcing)
  (exists (?P1 ?P2 ?T1 ?T2)
    (and
      (holdsDuring ?T1
        (spouse ?P1 ?P2))
      (holdsDuring ?T2
        (not (spouse ?P1 ?P2)))
      (finishes ?T1 ?D)
      (starts ?T2 ?D))))
```

Sumo includes a variety of domain ontologies

- communications,
- countries and regions,
- economy, finance,
- automobiles and engineering components,
- food,
- sports,
- shopping catalogs and Hotels,
- geography,
- government and Justice,
- language taxonomy,
- media and music,
- military (general, devices, processes, people),
- people and their emotions,
- physical elements,
- ...

Sumo website: http://www.adampease.org/OP/

Suggested Upper Merged Ontology (SUMO)

[Home | Browse | Download | Publications | Tools | What's New | About]

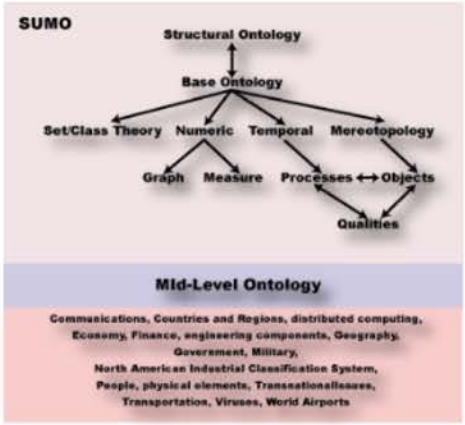
The Suggested Upper Merged Ontology (SUMO) and its domain ontologies form the largest formal public ontology in existence today. They are being used for research and applications in search, linguistics and reasoning. SUMO is the only formal ontology that has been mapped to all of the WordNet lexicon. SUMO is written in the SUO-KIF language. SUMO is free and owned by the IEEE. The ontologies that extend SUMO are available under GNU General Public License. Adam Pease is the Technical Editor of SUMO.

Find an English word and its corresponding formal term in SUMO:

English Word: table Noun Show

Features

- Mappings to all of WordNet
- Language generation templates for Hindi, Chinese, Italian, German, Czech and English
- Tools for browsing, editing, inferencing and NLP with SUMO are found in the SigmaKEE (source, api, docker image) and SigmaNLP (source, api, docker image) systems
- Largest free, formal ontology available, with ~25,000 terms and ~80,000 axioms when all domain ontologies are combined. These consist of SUMO itself, the MId-Level Ontology (MILO), and ontologies of Communications, Countries and regions, Distributed computing and User interfaces, Economy, Finance, Automobiles and Engineering components, Food, Dining, Sports, Shopping catalogs and Hotels, Geography, Government and Justice, Language taxonomy, Law, Media and Music, Military (General, Devices, Processes, People), North American Industrial Classification System, People and their Emotions, Physical elements, Transnational issues, Transportation and its Details, Viruses, Weather, world airports, and weapons of mass destruction. See also a large amount of instance content from DBPedia about people and the YAGO, project which includes millions of facts from Wikipedia merged with SUMO, and an initial merge of the Mondial geographical data with SUMO. The Open Biomedical Ontologies are lightly mapped to SUMO. Additional ontologies of terrorism are available on request.
- Richly axiomatized, not just a taxonomy. All terms are formally defined. Meanings are not dependent on a particular inference implementation.

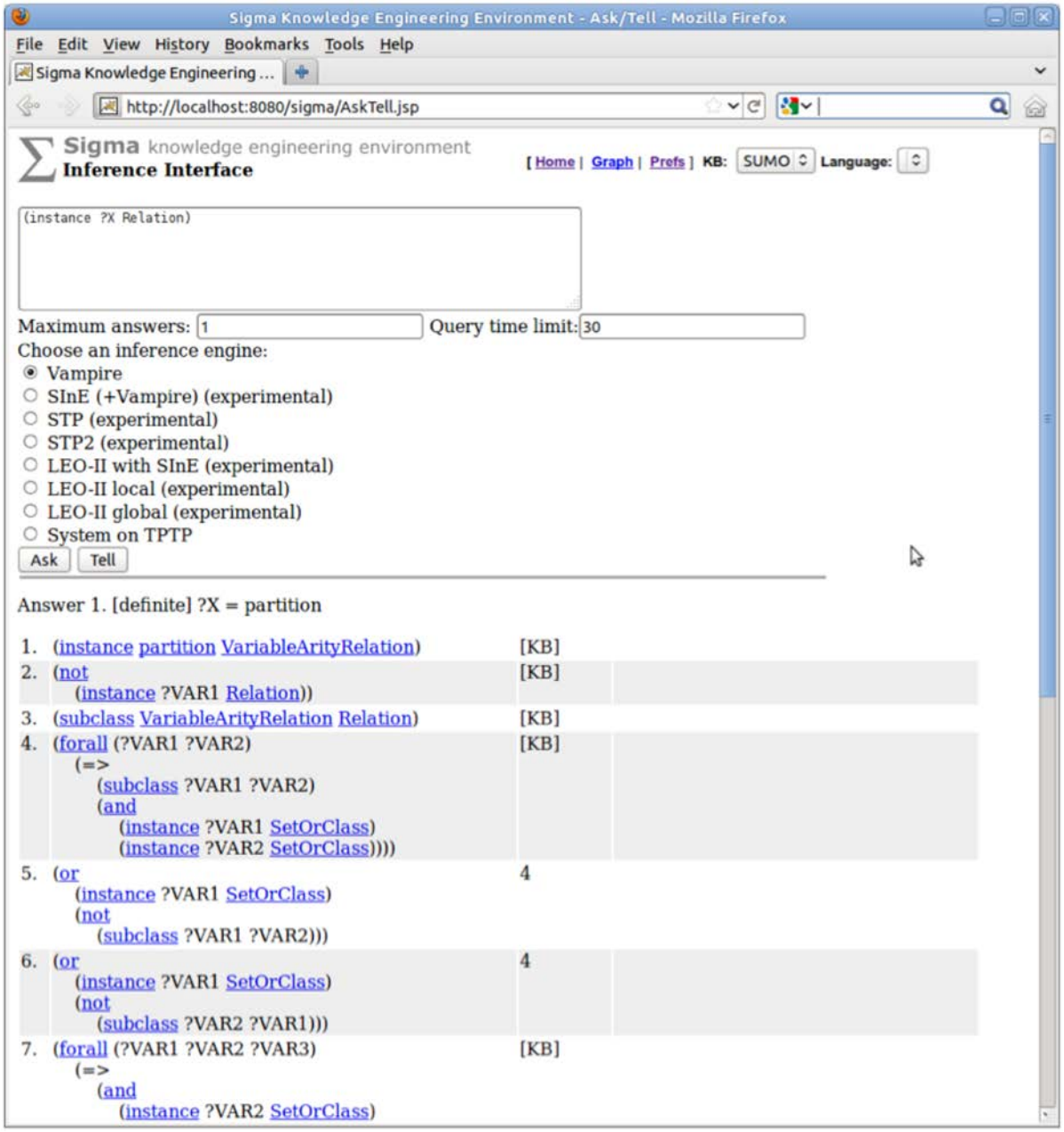


Documentation and other resources

- An introductory talk, **podcasts** and a blog about SUMO and Sigma
- Some good examples that show why simple search isn't enough and that we need inference to synthesize information from multiple sources
- look at historical versions of SUMO
- We ask that people working with SUMO cite our primary paper and book in any publications
- Conformance testing for SUMO
- Translation of SUMO into OWL and WordNet 3.0 in OWL
- WordNet in TPTP format
- Some thoughts on an ontology development process and ontology development pitfalls
- An introduction to resolution theorem proving
- **Frequently asked questions**
- A word sense frequency analysis based on WordNet SemCor that can be used for simple word sense disambiguation. This was done on WN 1.6 but could be rerun with the automatic remappings to 2.0
- A very big graph of the taxonomy in SUMO v 1.75 (too big to display in a browser, download and view in a drawing application)

Order a copy of the book "Ontology: A Practical Guide" for $25+tax/domestic US shipping (international customers should email me first to get shipping costs)

Sigma: An integrated development environment [Pease, Benzmüller, 2002]



What are the strengths and weaknesses of Sumo?

- Strengths
  - Open source initiative
  - Expressive representation and inference (SUO-KIF, Vampire)
  - Software resources (e.g., Sigma)
- Weaknesses
  - Small size compared to Cyc: 81K axioms (Cyc 7M assertions)
  - Incomplete content

There are other initiatives related to Cyc and Sumo

- ThoughtTreasure
  - A common sense knowledge base initiated by E. Mueller. Size: 27,000 concepts and 51,000 assertions (from 1993 to 2000) [Mueller, 1998]
- Upper ontologies
  - There a multiple proposals developed as upper ontologies (BFO, DOLCE, GFO, COSMO, etc.)

What are the strengths and weaknesses of manual acquisition of common sense knowledge?

- Strengths
  - The content of the knowledge base can grow in a more controlled way
  - Implicit common sense knowledge, which is not written in texts, can be identified and formulated by knowledge engineers
  - Expressive logic-based languages can be used with powerful inference methods
- Weaknesses
  - The effort required for manual acquisition of common sense knowledge is considerable
  - The current common sense knowledge bases are that are built manually are incomplete

Course “Common sense reasoning”.

Work citation in APA style:

> Molina, M. (2019). Common sense reasoning [Lecture slides]. OpenCourseWare, Universidad Politécnica de Madrid. Retrieved from http://ocw.upm.es/course