TableILP: Semi-Structured Reasoning for Answering Science Questions

Daniel Khashabi, Dan Roth (UIUC)
Tushar Khot, Ashish Sabharwal, Peter Clark, Oren Etzioni
(Allen Institute for Artificial Intelligence)

New Zealand shortest night

In ~~New York State~~, the ~~longest~~ period of ~~daylight~~ occurs during which month?

(A) June
(B) March
(C) December
(D) September

Premise: *a system that "understands" this phenomenon can correctly answer many variations!*

ILLINOIS UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

Semi-Structured Inference

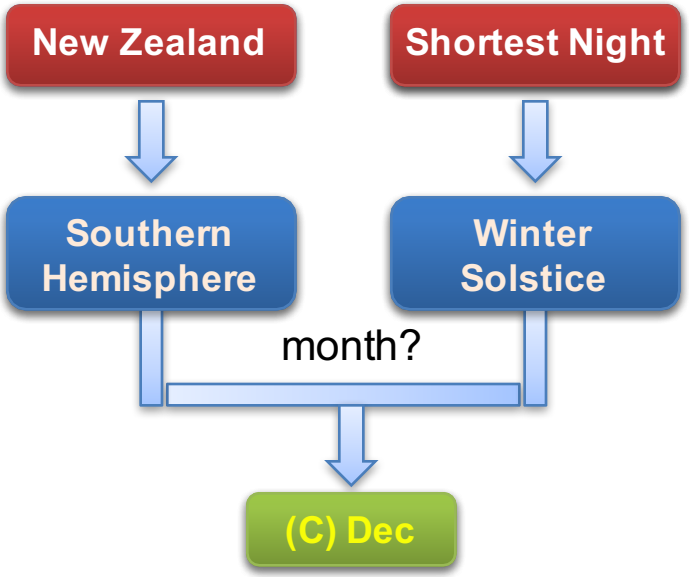
New Zealand shortest night

In ~~New York State~~, the ~~longest~~ period of ~~daylight~~ occurs during which month?
(A) June
(B) March
(C) December
(D) September

- Structured, Multi-Step Reasoning
  - science knowledge in small, manageable, swappable pieces: *regions, hemispheres, solstice*
  - Goal: overcome brittleness

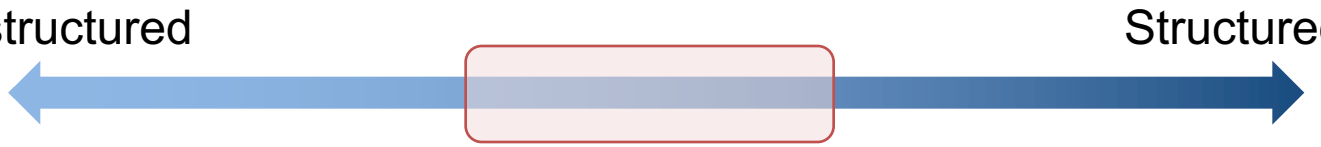
✓ principled approach, explainable answers
✓ robust to variations

How can we achieve this?



ILLINOIS UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
AI2 ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

Knowledge as Relational Tables



e.g., free form text from books, web

easy to acquire, difficult to reason with

Relational Tables with free form text

collections of recurring, related, science concepts

e.g., probabilistic first-order logic rules, ontologies

"easy" to reason with, difficult to acquire

Country	*Location*
France	north hemisphere
USA	north hemisphere
...	
Brazil	south hemisphere
Zambia	south hemisphere
...	

Hemisphere	*Orbital Event*	*Month*
northern	summer solstice	Jun
northern	winter solstice	Dec
northern	autumn equinox	Sep
...		
southern	summer solstice	Dec
southern	autumn equinox	Mar
...		

Energy, Forces, Adaptation, Phase Transition, Organ Function, Tools, Units, Evolution, …

Available at allenai.org

Simple structure, flexible content

- Can acquire knowledge in automated and semi-automated ways

ILLINOIS UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
AI2 ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

TableILP: Main Idea

Search for the best **Support Graph** connecting the Question to an Answer through Tables.

Q: In New York State, the longest period of daylight occurs during which month?

Cities, States, Countries	**Orbital Events:** **Geographical properties & Timing**
Potential Link: **Regions and Hemispheres**	

(A) December
(B) June
(C) March
(D) September

ILLINOIS UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
AI2 ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

TableILP: Main Idea

Search for the best **Support Graph** connecting the Question to an Answer through Tables.

Link this information to identify the best supported answer!

Q: In New York State, the longest period of daylight occurs during which month?

Subdivision	Country
New York State	USA
California	USA
Rio de Janeiro	Brazil
...	...

Orbital Event	Day Duration	Night Duration
Summer Solstice	Long	Short
Winter Solstice	Short	Long
....	....	...

Country	Hemisphere
United States	Northern
Canada	Northern
Brazil	Southern
.....	...

Hemisphere	Orbital Event	Month
North	Summer Solstice	June
North	Winter Solstice	December
South	Summer Solstice	December
South	Winter Solstice	June

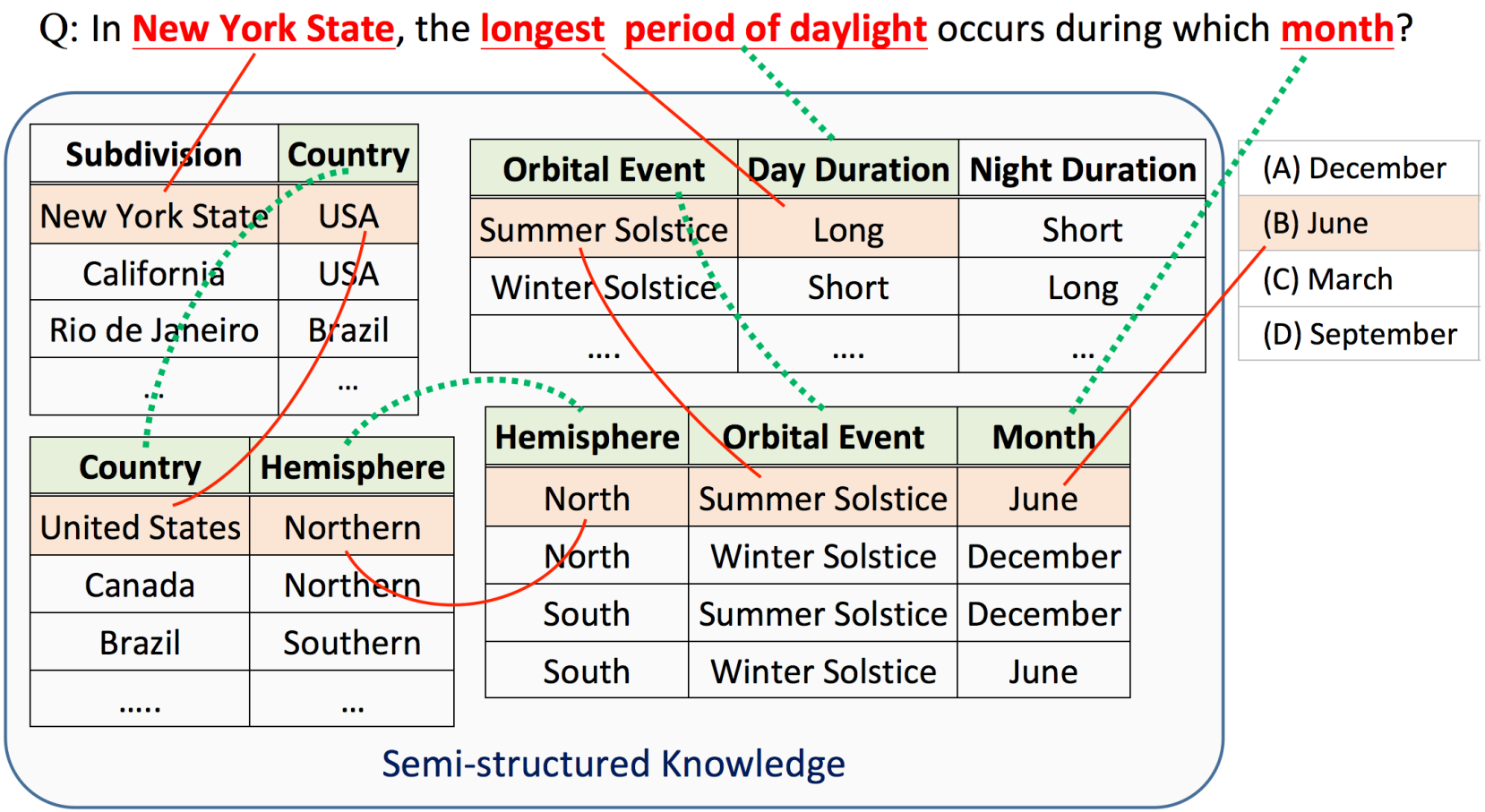
Semi-structured Knowledge

(A) December
(B) June
(C) March
(D) September

TableILP: Main Idea

Search for the best **Support Graph** connecting the Question to an Answer through Tables.

Link this information to identify the best supported answer!

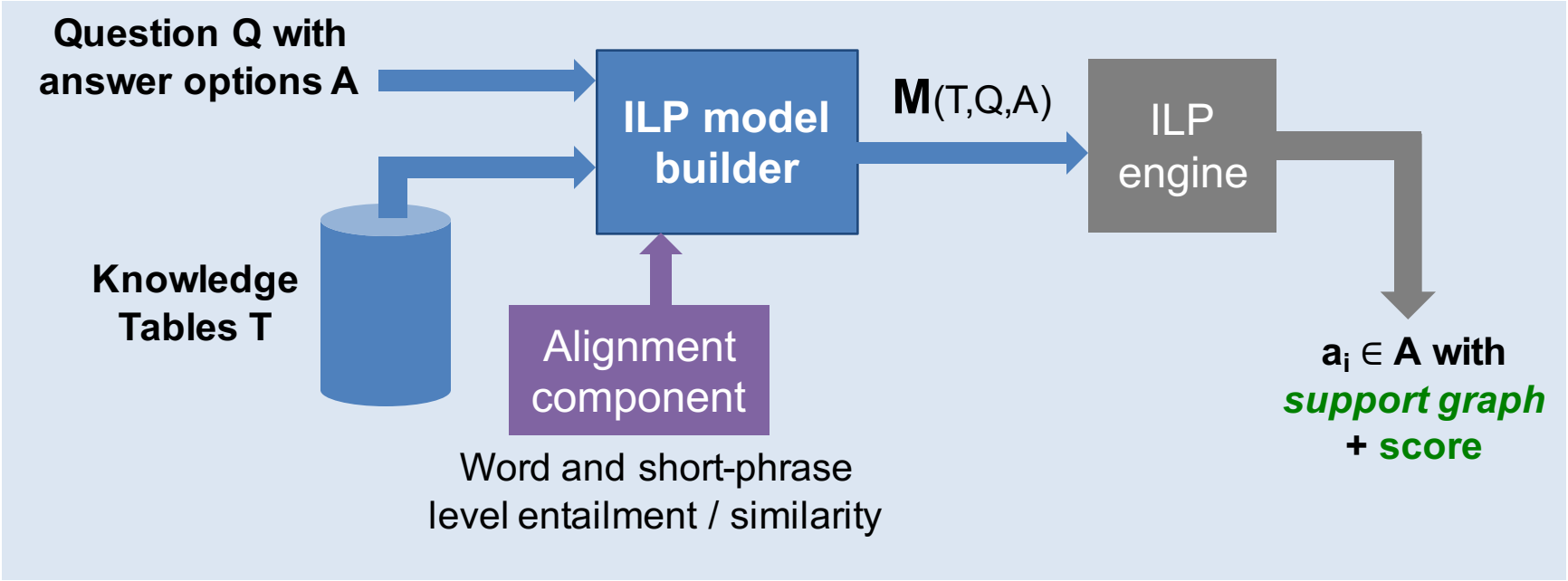


ILLINOIS UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

TableILP Solver: Overview

A discrete constrained **optimization** approach to QA for multiple-choice questions

- for each given question and candidate answers, we automatically generate a corresponding ILP objective and a set of constraints.



M(T,Q,A) →

$$\max \sum_i c_i x_i$$

$$\forall\, x_i \in \mathbb{N} \cup \{0\}$$

$$\begin{cases} \sum_i a_{1i} x_i \le b_1 \\ \dots \\ \sum_i a_{ki} x_i \le b_k \end{cases}$$

Optimization using Integer Linear Prog. formalism

ILLINOIS UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
AI2 ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

Approach: Integer Linear Program (ILP) Model

Goal: Design ILP constraints C and objective function F, s.t. maximizing F subject to C yields a "desirable" support graph

Variables define the space of "support graphs"

- Which nodes + edges between lexical units are active?

Objective Function: "better" support graphs = higher objective value

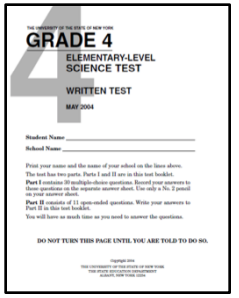
- Reward active units, high lexical match links, column header match, …
- Penalize spurious overuse of frequently occurring terms

Constraints

- ~50 high-level constraints
  - Basic Lookup, Parallel Evidence, Evidence Chaining, Semantic Relation Matching
- Examples: connectedness, question coverage, appropriate table use

Evaluation

- **4th Grade NY Regents Science Exam**
  - Focus on non-diagram multiple-choice (4-way)
  - 129 questions in completely unseen Test set
    - 6 years of exams; 95% C.I. = 9%
  - Score: 1 point per question (1/k for k-way tie including correct answer)


GRADE 4
ELEMENTARY-LEVEL SCIENCE TEST
WRITTEN TEST
MAY 2004

Student Name _____
School Name _____

Print your name and the name of your school on the lines above.
The test has two parts. Parts I and II are in this test booklet.
Part I contains 20 multiple-choice questions. Record your answers to these questions on the separate answer sheet. Use only a No. 2 pencil on your answer sheet.
Part II consists of 11 open-ended questions. Write your answers to Part II in this test booklet.
You will have as much time as you need to answer the questions.

DO NOT TURN THIS PAGE UNTIL YOU ARE TOLD TO DO SO.

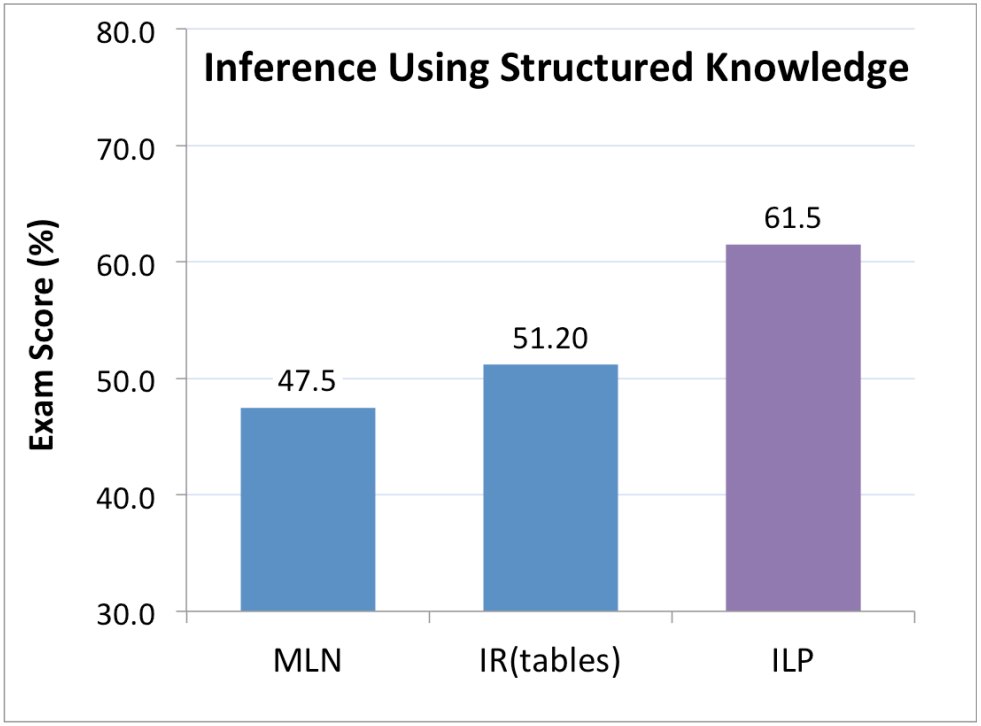
Copyright 2004
THE UNIVERSITY OF THE STATE OF NEW YORK
THE STATE EDUCATION DEPARTMENT
ALBANY, NEW YORK 12234

Available at allenai.org

- **Baselines:**
  - IR Solver: Information Retrieval using Lucene search
    - Using 280 GB of plain text (50B tokens) "waterloo" corpus [AAAI, 2015]
    - IR Solver(tables): Using same tables as TableILP
  - PMI Solver: Statistical correlation using pointwise mutual info.
    - Using 280 GB of plain text (50B tokens) "waterloo" corpus [AAAI, 2015]
  - MLN: Markov Logic Network, a structured prediction model
    - Using rules from 80K sentences [EMNLP, 2015]

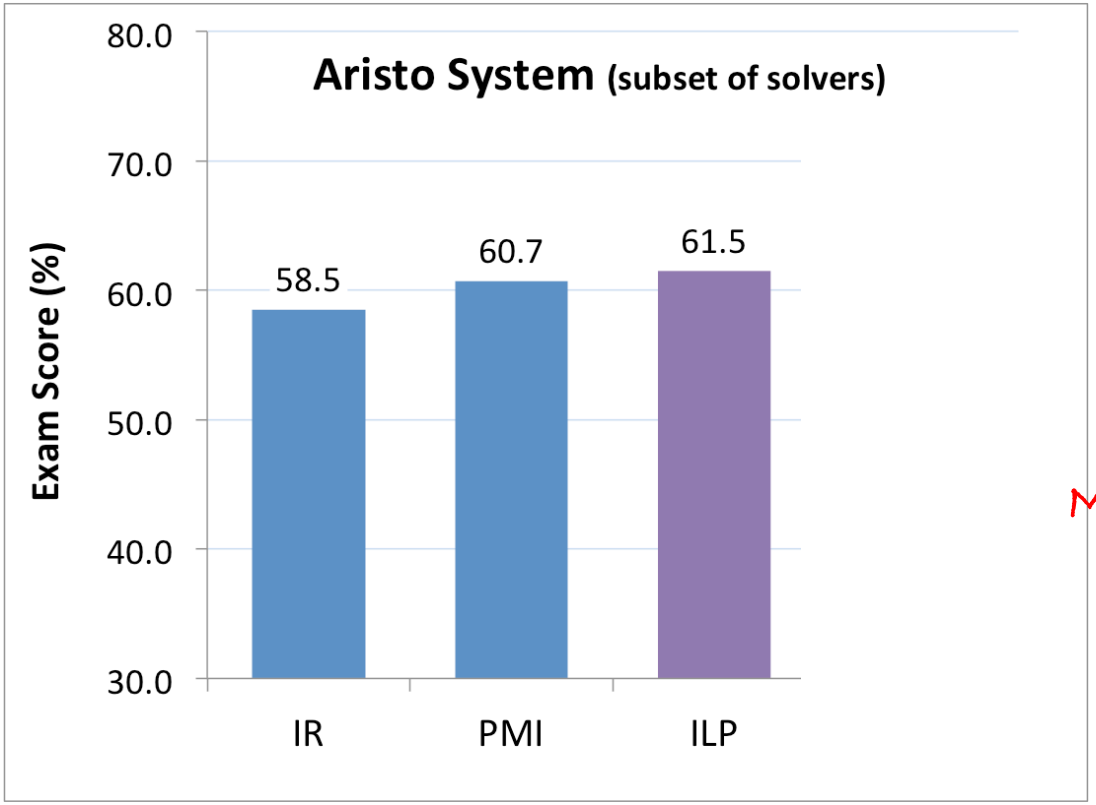
Results: Same Knowledge

TableILP is substantially better than IR & MLN, when given knowledge derived from the same, domain-targeted sources



ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

Results



More experiments in the paper!

Ensemble performs 8-10% higher than IR baselines

Simple logistic regression. Features: [Clark et al, AAAI-2016]

- 4 from each solver's score
- 11 from TableILP's support graph (#rows, weakest edge, …)

ILLINOIS UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

Conclusions

- **TableILP: Semi-structured reasoning** can be very effective
  - Beyond IR
  - Just starting to scratch the surface!
  - Code: https://github.com/allenai/tableilp

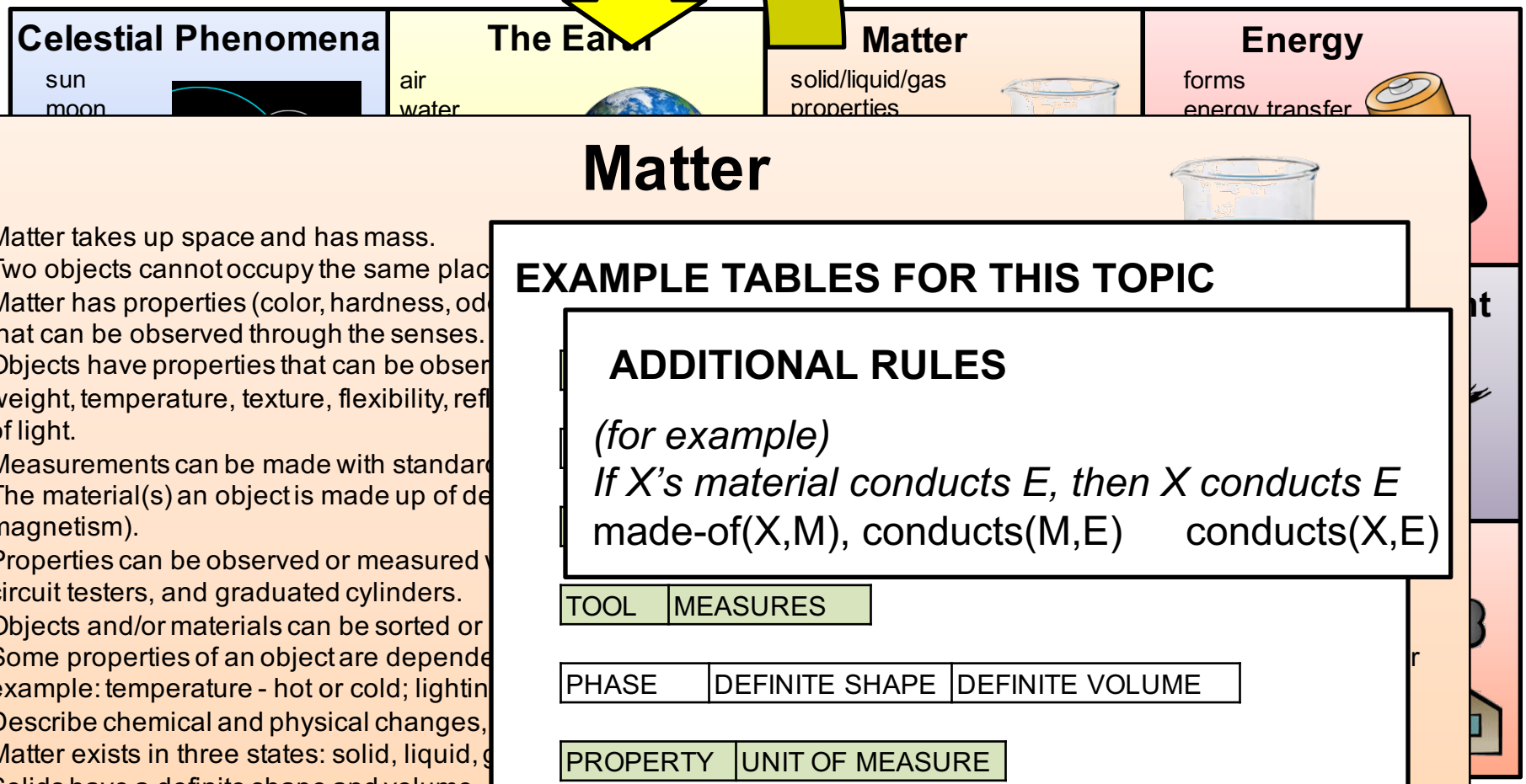
- Ongoing efforts + future extensions
  - Scaling up to medium/large scale KB
  - Automated parameter tuning / learning
  - Improved semantics (better question interpretation, negations, …

EXTRA SLIDES

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
AI2 ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

Knowledge as Relational Tables

- The Knowledge Atlas [illegible] k[illegible]sections



Matter

Matter takes up space and has mass.
Two objects cannot occupy the same plac
Matter has properties (color, hardness, od
that can be observed through the senses.
Objects have properties that can be obser
weight, temperature, texture, flexibility, refl
of light.
Measurements can be made with standar
The material(s) an object is made up of de
magnetism).
Properties can be observed or measured
circuit testers, and graduated cylinders.
Objects and/or materials can be sorted or
Some properties of an object are depende
example: temperature - hot or cold; lightin
Describe chemical and physical changes,
Matter exists in three states: solid, liquid, g
Solids have a definite shape and volume
Liquids do not have a definite shape but h
Gases do not hold their shape or volume
Temperature can affect the state of matter of a substance.
Changes in the properties or materials of objects can be observed and described.

EXAMPLE TABLES FOR THIS TOPIC

ADDITIONAL RULES

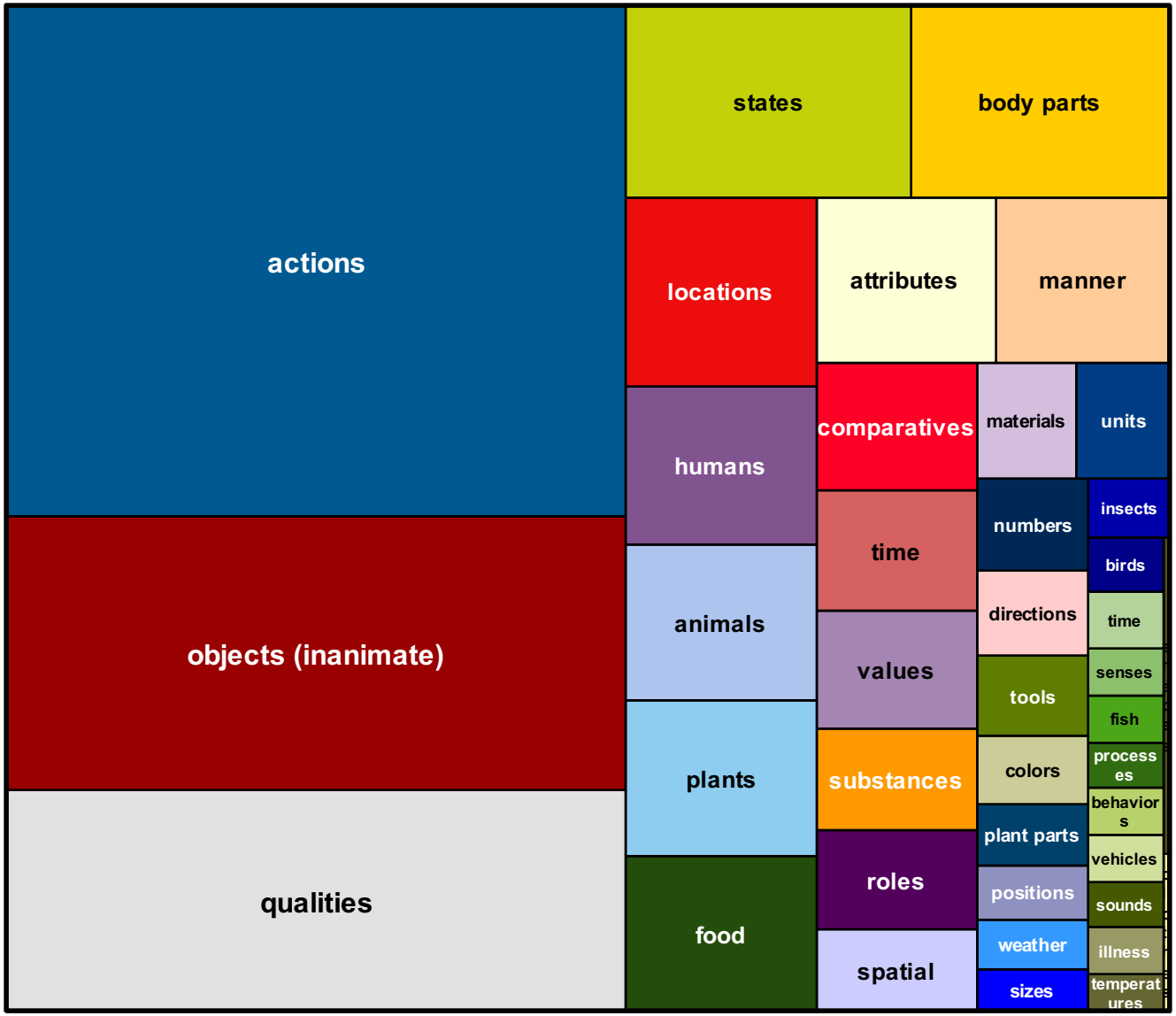
(for example)
If X's material conducts E, then X conducts E
made-of(X,M), conducts(M,E) conducts(X,E)

TOOL	MEASURES

PHASE	DEFINITE SHAPE	DEFINITE VOLUME

PROPERTY	UNIT OF MEASURE

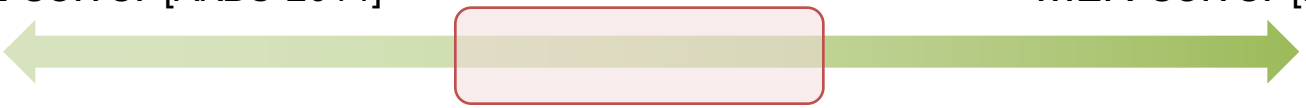
Relation Involving Which Objects?



Grouping of ~2500 key terms related to 4th grade science

Semi-Structured Inference: Challenge #2

Reasoning: effective, controllable, scalable



forward chaining of logic rules

Pros:
easy to understand behavior (state space)

Cons:
focuses on *how* to search rather than *what* to look for

Integer Linear Programming (ILP) framework

constraints and preferences, industrial-strength solvers

approx. inference with probabilistic first-order logic

Pros:
"natural" fit, high-level specification

Cons:
inefficient, difficult to control, brittle with noisy input

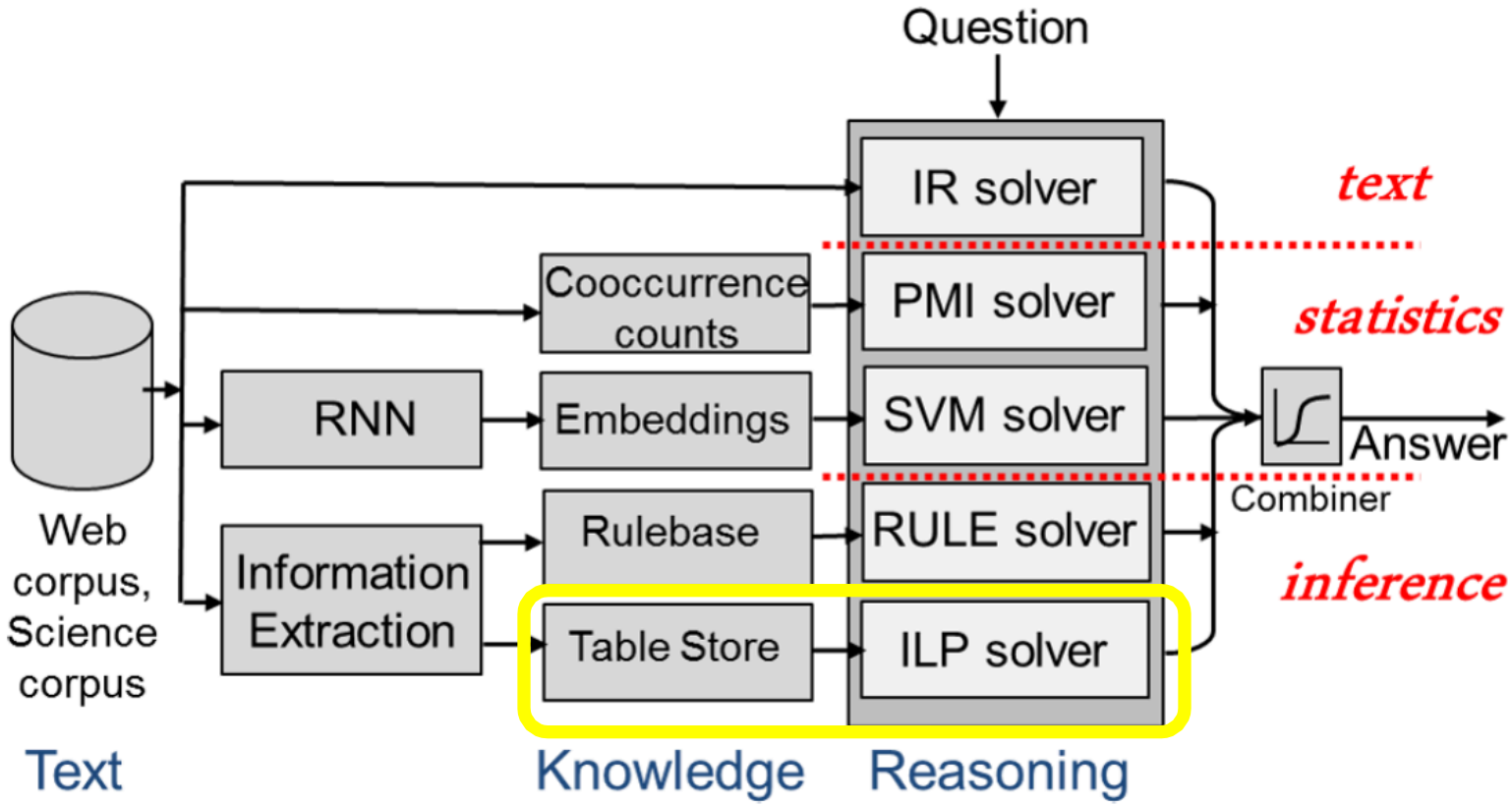
Evaluation: Ablation Study

- Key components of the TableILP system contribute substantially to the eventual score

Solver	Test Score (%)
TableILP	61.5
No Multiple Row Inference	51.0
No Relation Matching	55.6
No Open IE Tables	52.3
No Lexical Entailment	50.5

Aristo: Ensemble Approach

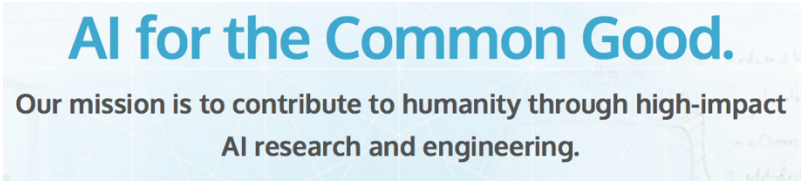
[AAAI-2016]



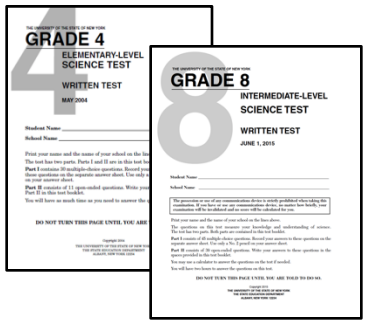
ILLINOIS UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

Three Takeaways

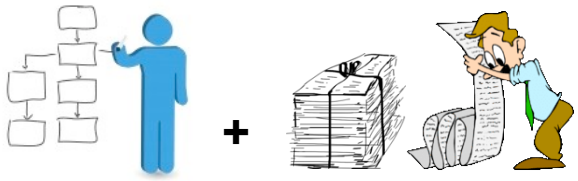
1. **AI2**: exciting place for cutting-edge AI research and engineering!



2. **Standardized exams** (science, math, …): great test beds for pushing AI & assessing progress
   - Super-interesting, challenging, measurable
   - Just starting to scratch the surface!

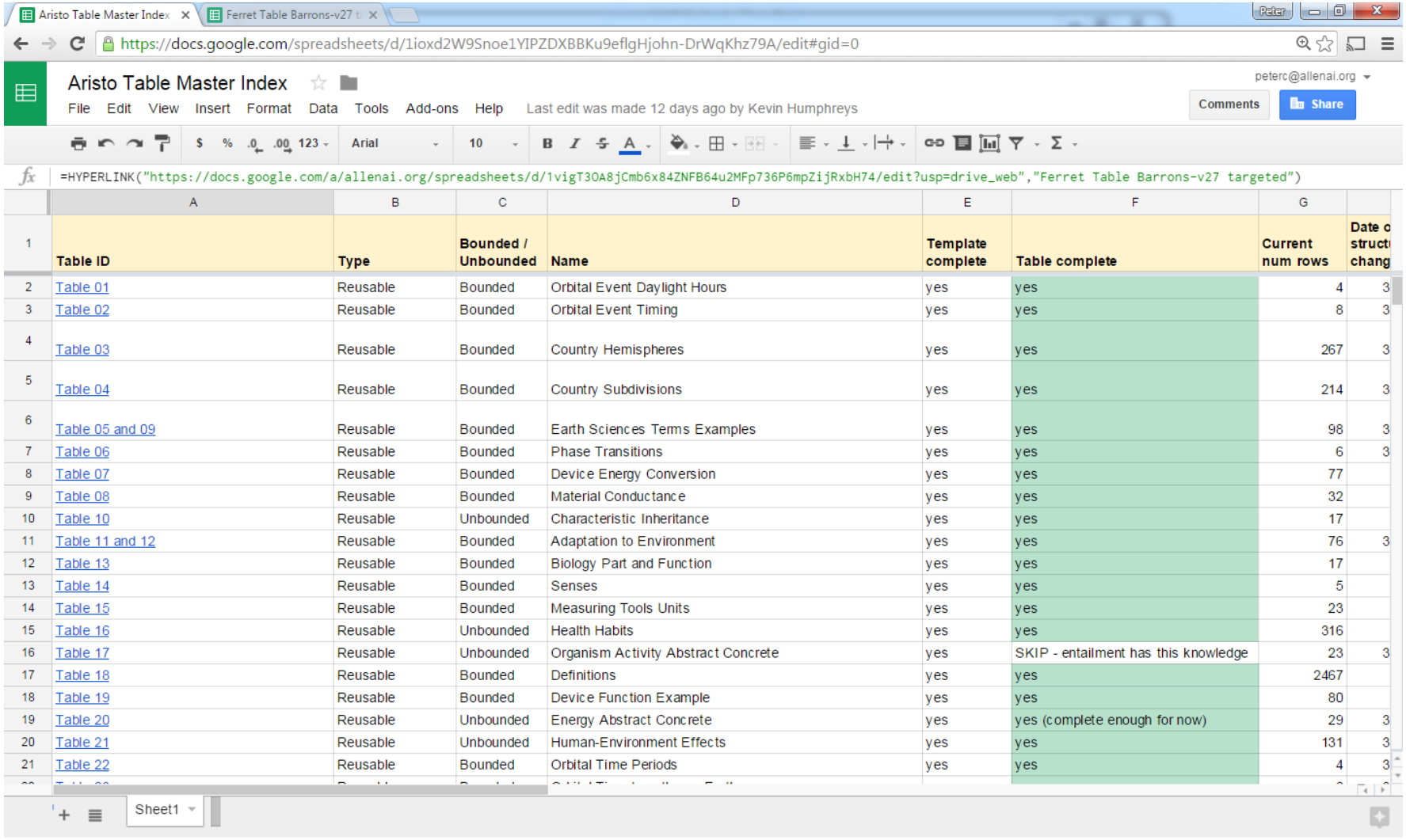


1. **Semi-structured inference** can be very effective & robust on these tests
   - Goes beyond factoid-style QA
   - Complementary to IR



Aristo's Tablestore

- ~85 tables, ~10k rows, ~30k cells
- Defined with respect to questions, study guides, syllabus



Aristo Table Master Index | Ferret Table Barrons-v27

https://docs.google.com/spreadsheets/d/1ioxd2W9Snoe1YIPZDXBBKu9eflgHjohn-DrWqKhz79A/edit#gid=0

Aristo Table Master Index

File Edit View Insert Format Data Tools Add-ons Help Last edit was made 12 days ago by Kevin Humphreys

peterc@allenai.org

Comments Share

=HYPERLINK("https://docs.google.com/a/allenai.org/spreadsheets/d/1vigT3OA8jCmb6x84ZNFB64u2MFp736P6mpZijRxbH74/edit?usp=drive_web","Ferret Table Barrons-v27 targeted")

	A	B	C	D	E	F	G	
1	Table ID	Type	Bounded / Unbounded	Name	Template complete	Table complete	Current num rows	Date o struct chang
2	Table 01	Reusable	Bounded	Orbital Event Daylight Hours	yes	yes	4	3
3	Table 02	Reusable	Bounded	Orbital Event Timing	yes	yes	8	3
4	Table 03	Reusable	Bounded	Country Hemispheres	yes	yes	267	3
5	Table 04	Reusable	Bounded	Country Subdivisions	yes	yes	214	3
6	Table 05 and 09	Reusable	Bounded	Earth Sciences Terms Examples	yes	yes	98	3
7	Table 06	Reusable	Bounded	Phase Transitions	yes	yes	6	3
8	Table 07	Reusable	Bounded	Device Energy Conversion	yes	yes	77	
9	Table 08	Reusable	Bounded	Material Conductance	yes	yes	32	
10	Table 10	Reusable	Unbounded	Characteristic Inheritance	yes	yes	17	
11	Table 11 and 12	Reusable	Bounded	Adaptation to Environment	yes	yes	76	3
12	Table 13	Reusable	Bounded	Biology Part and Function	yes	yes	17	
13	Table 14	Reusable	Bounded	Senses	yes	yes	5	
14	Table 15	Reusable	Bounded	Measuring Tools Units	yes	yes	23	
15	Table 16	Reusable	Unbounded	Health Habits	yes	yes	316	
16	Table 17	Reusable	Unbounded	Organism Activity Abstract Concrete	yes	SKIP - entailment has this knowledge	23	3
17	Table 18	Reusable	Bounded	Definitions	yes	yes	2467	
18	Table 19	Reusable	Bounded	Device Function Example	yes	yes	80	
19	Table 20	Reusable	Unbounded	Energy Abstract Concrete	yes	yes (complete enough for now)	29	3
20	Table 21	Reusable	Unbounded	Human-Environment Effects	yes	yes	131	3
21	Table 22	Reusable	Bounded	Orbital Time Periods	yes	yes	4	3

Sheet1

ILP Complexity, Scalability

- ~50 high-level constraints

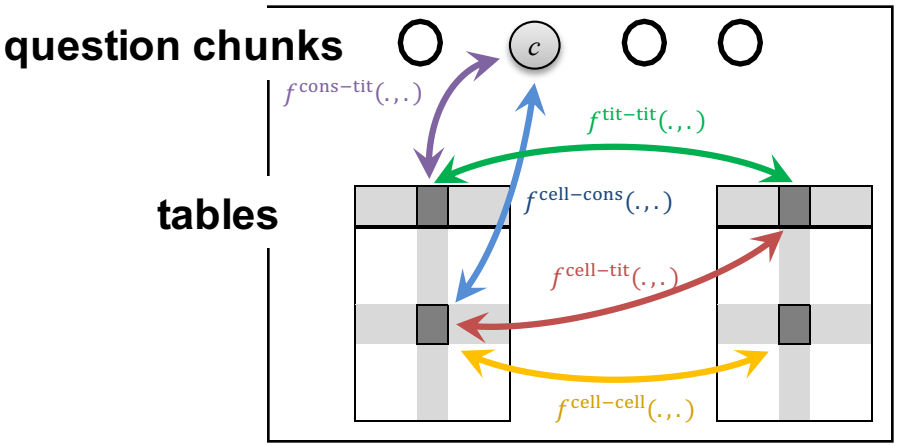
Category	Quantity	Average
ILP complexity	#variables	1043.8
	#constraints	4417.8
	#LP iterations	1348.9
Knowledge use	#rows	2.3
	#tables	1.3
Timing stats	model creation	1.9 sec
	solving the ILP	2.1 sec

- Speed: **4 sec** per question, reasoning over 140 rows across 7 tables
  - Contrast: **17 sec for MLN using only 1 rule** per answer option!
  - Commercial ILP engines (Gurobi, Cplex) much faster than SCIP

ILP Model

Operates on lexical units of alignment

- cells + headers of tables T
- question chunks Q
- answer options A



~50 high level constraints + preferences

Variables define the space of "support graphs" connecting Q, A, T

- Which nodes + edges between lexical units are active?

Objective Function: "better" support graphs = higher objective value

- Reward active units, high lexical match links, column header match, …
- WH-term boost (which form of energy), science-term boost (evaporation)
- Penalize spurious overuse of frequently occurring terms

ILP Model: Constraints

Dual goal: scalability, consider only meaningful support graphs

- **Structural Constraints**
  - Meaningful proof structures
    - connectedness, question coverage, appropriate table use
    - parallel evidence => identical multi-row activity signature
  - Simplicity appropriate for 4^{th} / 8^{th} grade

- **Semantic Constraints**
  - Chaining => table joins between semantically similar column pairs
  - Relation matching (ruler measures length, change from water to liquid)

- **Table Relevance Ranking**
  - TF-IDF scoring to identify top N relevant tables

Assessing Brittleness: Question Perturbation

How robust are approaches to simple question perturbations *that would typically make the question easier for a human*?

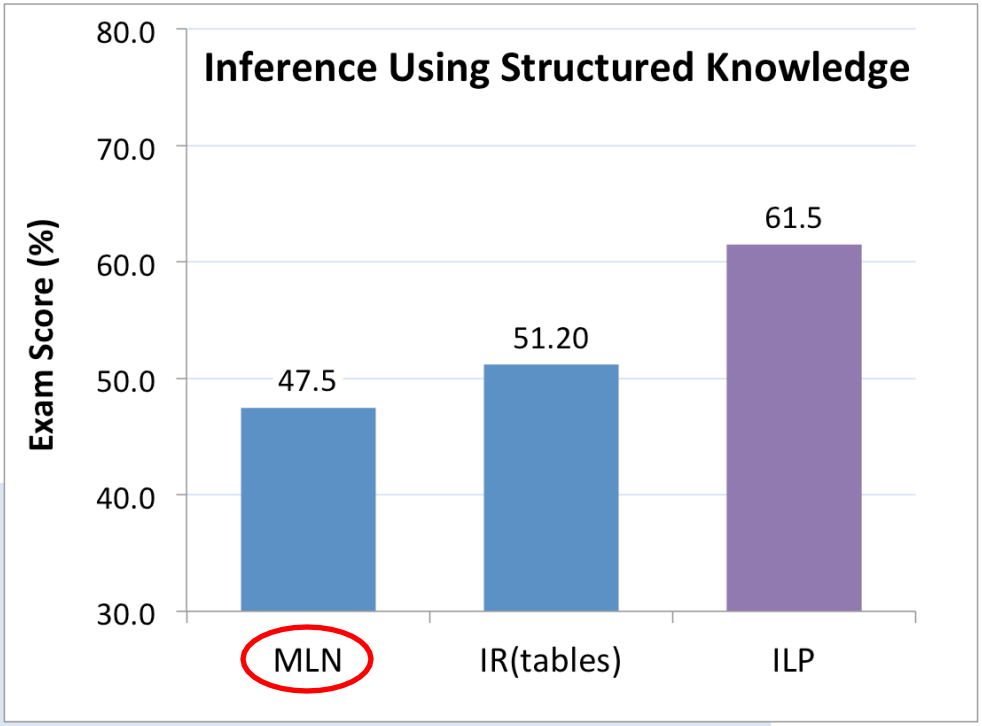
- E.g., Replace incorrect answers with arbitrary co-occurring terms

> In New York State, the longest period of daylight occurs during which month?
> (A) *eastern* (B) June (C) *history* (D) *years*

Solver	Original Score (%)	% Drop with Perturbation	
		absolute	relative
IR	70.7	13.8	19.5
PMI	73.6	24.4	33.2
TableILP	85.0	**10.5**	**12.3**

Results: Exploiting Structured Knowledge

TableILP is substantially better than IR & MLN, when given knowledge derived from the same, domain-targeted sources



[EMNLP-2015]

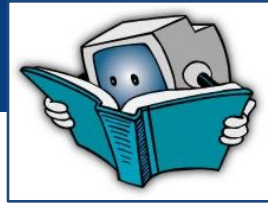
Best of 3 **MLN approaches**:

A. First-order rules "as is"
   - Convenient, natural
   - Slow, despite a few tricks

B. Entity Resolution based MLN
   - Probabilistic "SameAs" predicate
   - Much faster, but brittle – low recall

C. Customized MLN: controlled search for valid reasoning chains
   - More controllable, more robust, more scalable (but still very limited)

Standardized Tests as an AI Challenge



Build AI systems that demonstrate human-like intelligence by passing standardized science exams <u>as written</u>

Many challenges: broad knowledge (general and scientific), question interpretation, reasoning at the right level of granularity, …

Which physical structure would best help a bear to **survive a winter** in New York State?
(A) big ears (B) black nose (C**) thick fur** (D) brown eyes

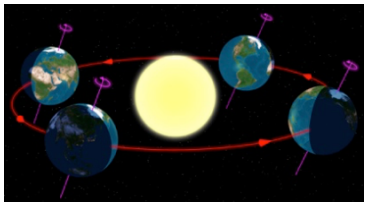
Two Approaches to Question Answering

New Zealand shortest night

In ~~New York State~~, the ~~longest~~ period of ~~daylight~~ occurs during which month?
(A) June
(B) March
(C) December
(D) September

Premise: *a system that "understands" this phenomenon can correctly answer many variations!*

- **Sophisticated physics model** of planetary movement
  - ✓ powerful model, would enable complex reasoning
  - × difficult to implement, scale up, or learn automatically



- **Information retrieval** / statistical association
  - ✓ easy, generalizes well, often effective
  - × limited to simple reasoning
  - × expects answers explicitly written somewhere