



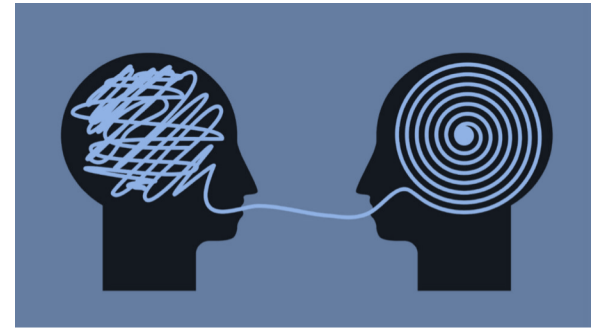
In Pursuit of the Holy Grail of Natural Language Understanding: Past, Present, and Future

September 2019

Daniel Khashabi

My Background

- Recently finished PhD (UPenn)
- Been working with Allen Institute for AI (AI2)
 - On-and-off since 2015, and full-time since this August
- Research theme:
 - Artificial Intelligence, through the lens of **natural language** understanding
 - Not specified:
 - What **solution** we use to achieve this goal
- Not a CP person!



- Founded in 2014 by Paul Allen (Microsoft co-founder)
 - Non-profit research organization
 - Mission: contribute to humanity through high-impact AI research and engineering



Oren Etzioni
CEO



AI for the Common Good.

Our mission is to contribute to humanity through high-impact AI research and engineering.



Project Aristo (2014-2019)

- **Vision: The Knowledgeable Machine**

- Large volumes of general and scientific knowledge, stored in a "computable" form, supporting reasoning and explanation

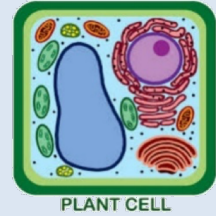
- **Measurable goal:**

- Pass elementary-school **science exams** as written
 - Currently 4th grade and 8th grade exams
- *Credit goes to dozens of researchers who have contributed to this project.*

Example Question: Reasoning by Chaining

The cell structure that makes a plant cell more rigid than an animal cell is the

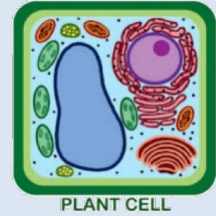
- (A) cell membrane.
- (B) cytoplasm.
- (C) cell wall.
- (D) ribosome.



Example Question: Reasoning by Chaining

The cell structure that makes a plant cell more rigid than an animal cell is the

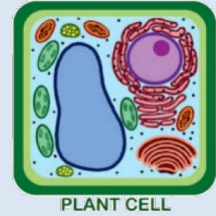
- (A) cell membrane.
- (B) cytoplasm.
- (C) cell wall.
- (D) ribosome.



Plants use cellulose for their cell walls.

Example Question: Reasoning by Chaining

The cell structure that makes a **plant cell** more rigid than an animal cell is the
(A) cell membrane.
(B) cytoplasm.
(C) **cell wall.**
(D) ribosome.

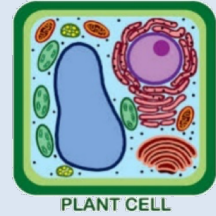


Plants use **cellulose** for their **cell walls.**

All fibers are assembled of chains of **cellulose molecules**, arranged as **a rigid structure.**

Example Question: Reasoning by Chaining

The cell structure that makes a plant cell more rigid than an animal cell is the
(A) cell membrane.
(B) cytoplasm.
(C) cell wall.
(D) ribosome.



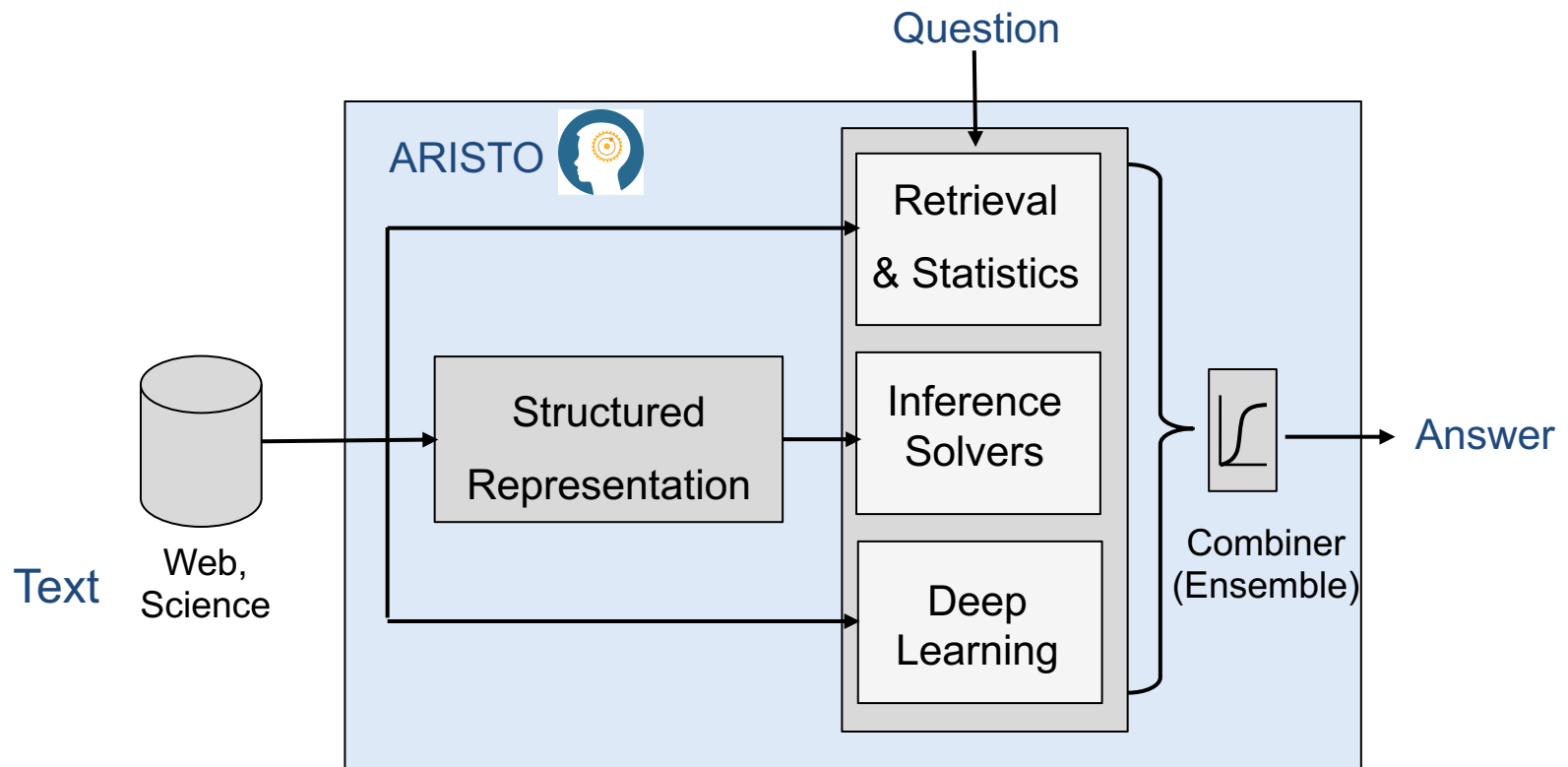
Plants use cellulose for their cell walls.

All fibers are assembled of chains of cellulose molecules, arranged as a rigid structure.

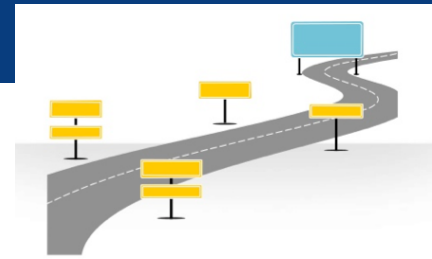


Aristo: an over-simplified overview

- An ensemble architecture
 - To deal with questions with a variety of difficulty

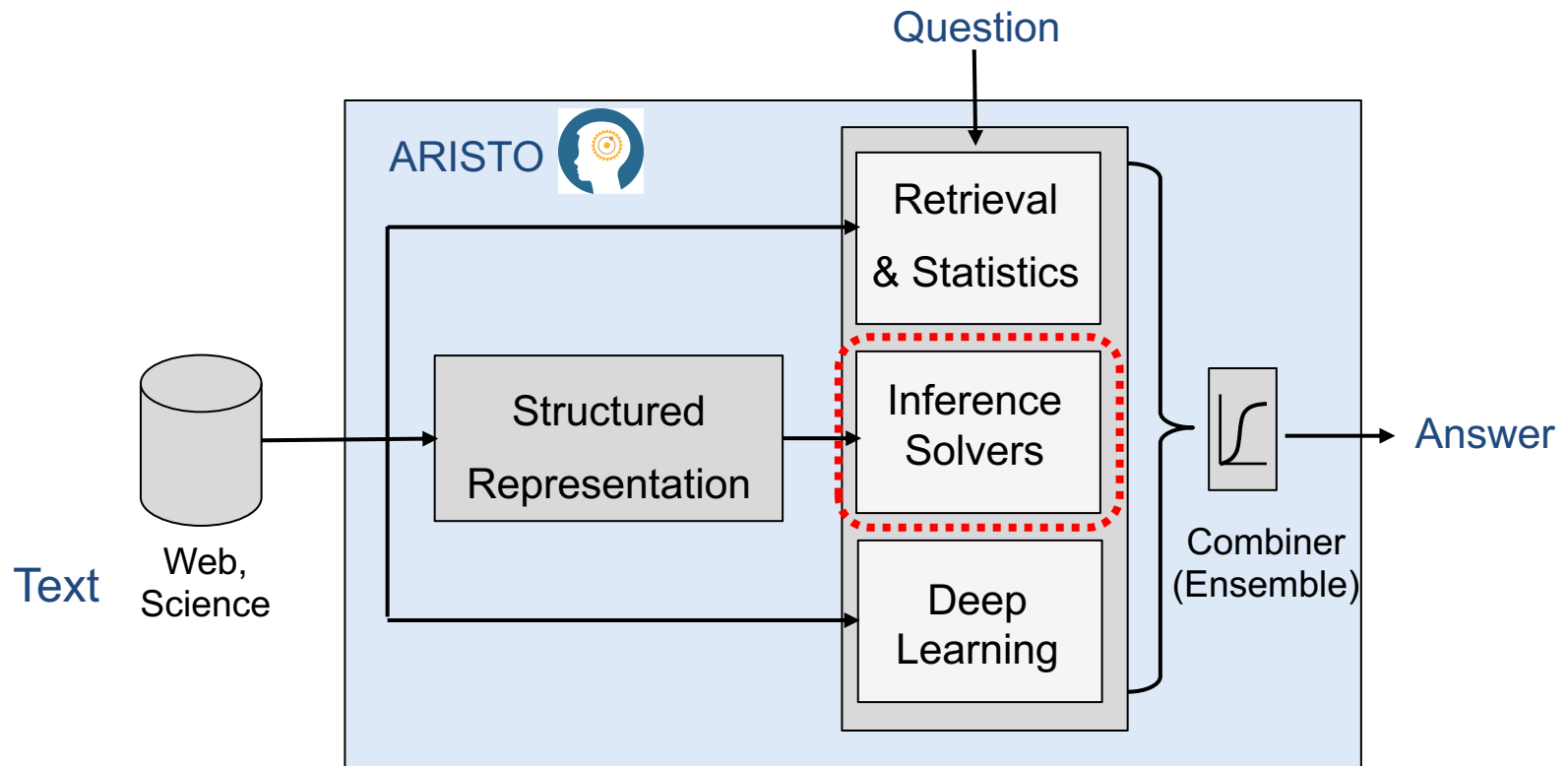


Road map



I. Aristo Inference Solvers

II. Beyond inference Modules



Aristo Inference Modules

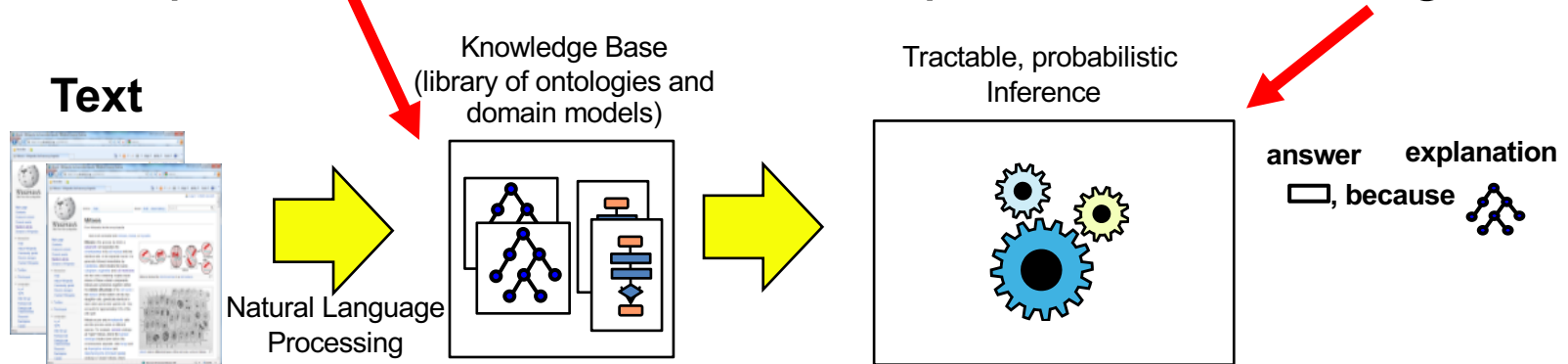
Khashabi et al, AAIL-2018

Khot et al, ACL-2017

Khashabi et al, IJCAI-2016

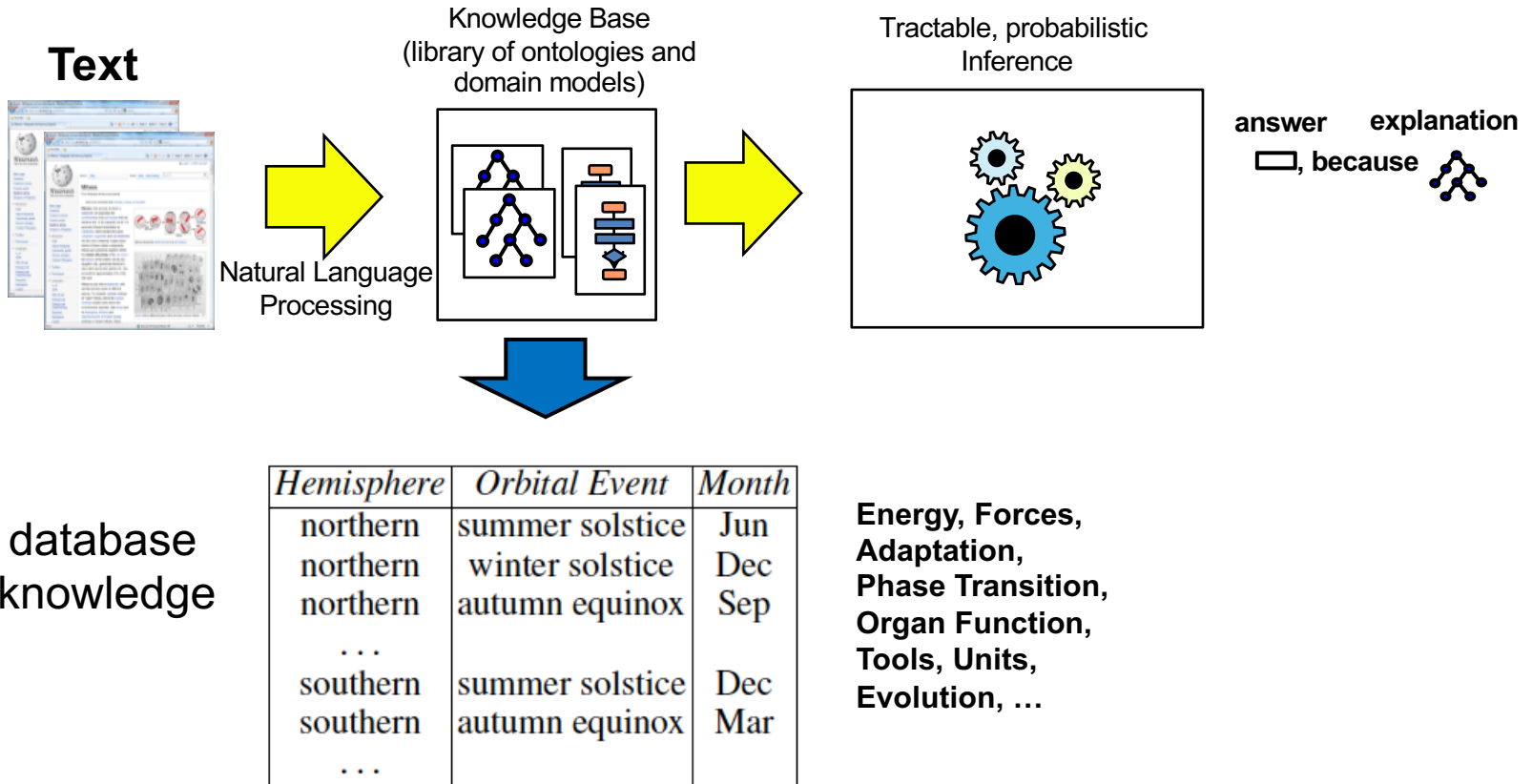
Aristo Inference Modules

- General structure of **inference frameworks**
 - Text & questions converted to an intermediate representation
 - A “probabilistic” inference for question-answering



- Three key scientific questions:
 - How do we identify and represent the **meaning** of text?
 - What mechanisms support **robust reasoning** with incomplete and incorrect knowledge?

Knowledge in Tabular Representation



Simple structure, flexible content

- Can acquire knowledge in automated and semi-automated ways
- No Cyc-like claims of “completeness” or “adequacy”

[Dalvi et al, 2016]

TableLP: Main Idea

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

Cities, States, Countries	Orbital Events: Geographical properties & Timing	(A) December
<i>Potential Link:</i> Regions and Hemispheres		(B) June
		(C) March
		(D) September

TableLP: 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?

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

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

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

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

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

Semi-structured Knowledge

TableLP: 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
...	...

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

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

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

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

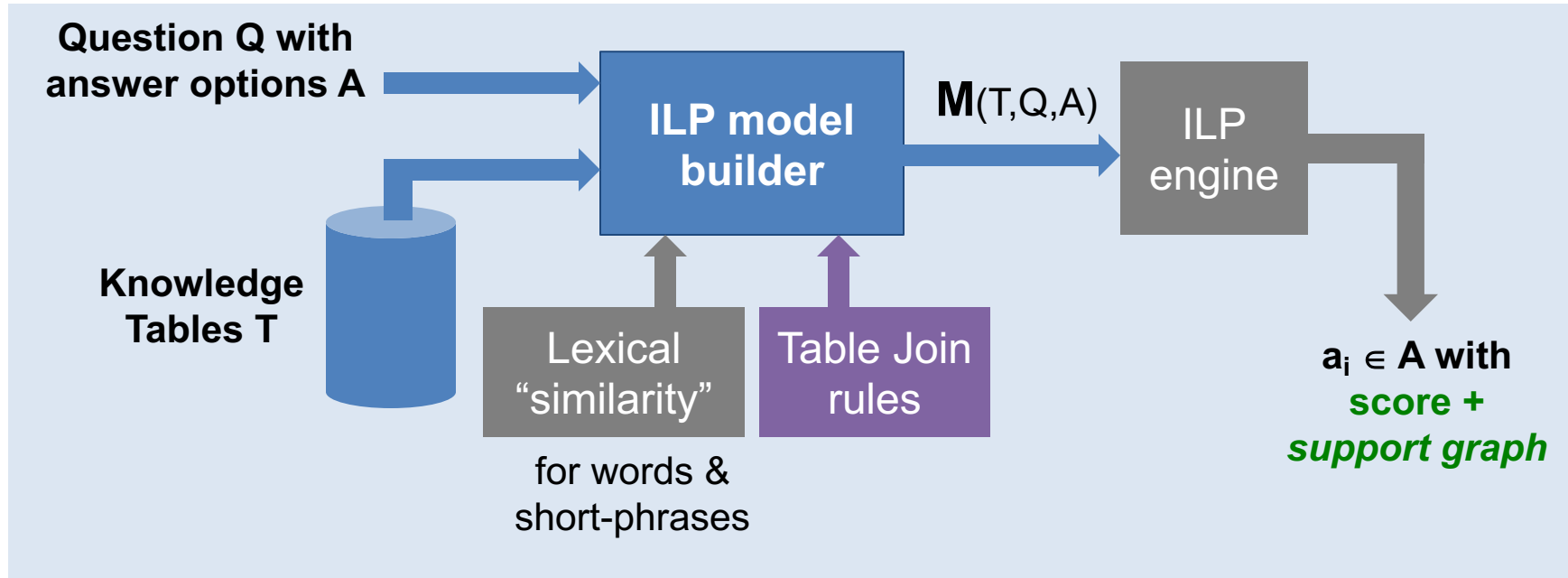
Semi-structured Knowledge

Abductive reasoning

[Peirce, 1883]

TableLP Solver: Overview

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



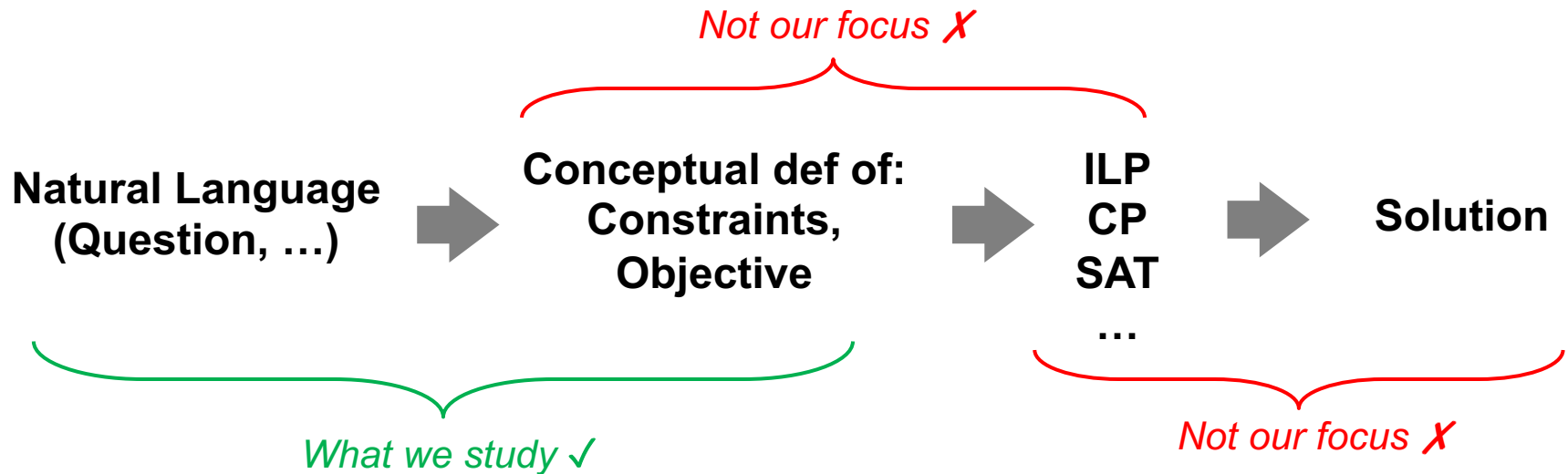
$M(G, Q, A) \rightarrow$

$$\begin{aligned} & \max \sum_i c_i x_i \\ & \forall x_i \in \mathbb{N} \cup \{0\} \end{aligned} \quad \left\{ \begin{array}{l} \sum_i a_{1i} x_i \leq b_1 \\ \dots \\ \sum_i a_{ki} x_i \leq b_k \end{array} \right.$$

Optimization using Integer Linear Program (ILP) formalism

ILP Model: Design Challenges

Goal: Design ILP objective function, s.t. maximizing it subject to the constraints yields a “desirable” support graph



ILP Model: Design Challenges

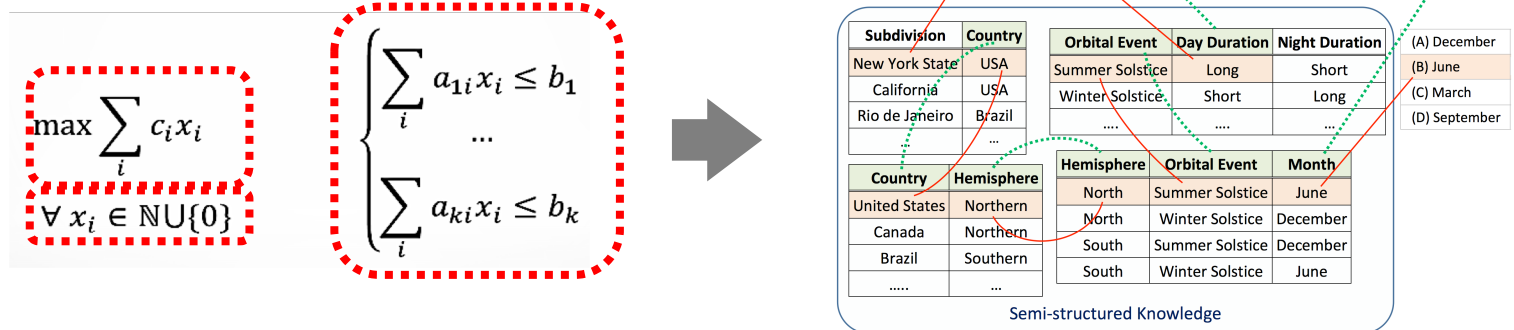
Goal: Design ILP objective function, s.t. maximizing it subject to the constraints yields a “desirable” support graph

Not so straightforward!

$$\begin{array}{ll} \max \sum_i c_i x_i & \left\{ \begin{array}{l} \sum_i a_{1i} x_i \leq b_1 \\ \dots \\ \sum_i a_{ki} x_i \leq b_k \end{array} \right. \\ \forall x_i \in \mathbb{N} \cup \{0\} & \end{array}$$

- Many possible “proof structures”
- Imperfect lexical “similarity” blackbox
- Partial or missing knowledge
- Question logic (negation, conjunction, comparison)
- Scalability of ILP solvers
- ...

ILP Model: Some Details



ILP Model: Some Details

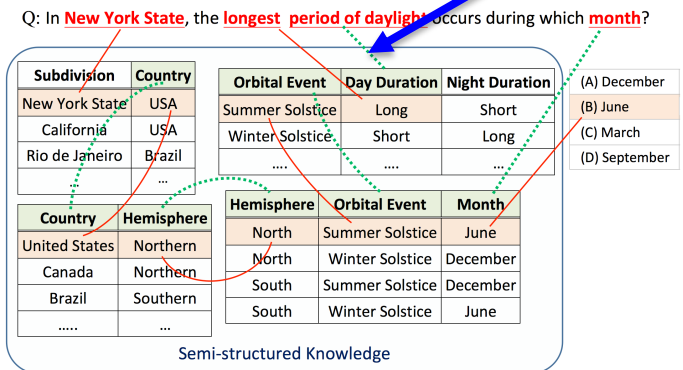
Variables define the space of “support graphs”:

- Each variable corresponds to to a node or edge.
- $x=1$ iff nodes / edges are part of the semantic graph.

$$\begin{aligned} \max \sum_i c_i x_i \\ \left\{ \begin{array}{l} \sum_i a_{1i} x_i \leq b_1 \\ \dots \\ \sum_i a_{ki} x_i \leq b_k \end{array} \right. \\ \forall x_i \in \mathbb{N} \cup \{0\} \end{aligned}$$



Support Graph



ILP Model: Some Details

Objective Function:

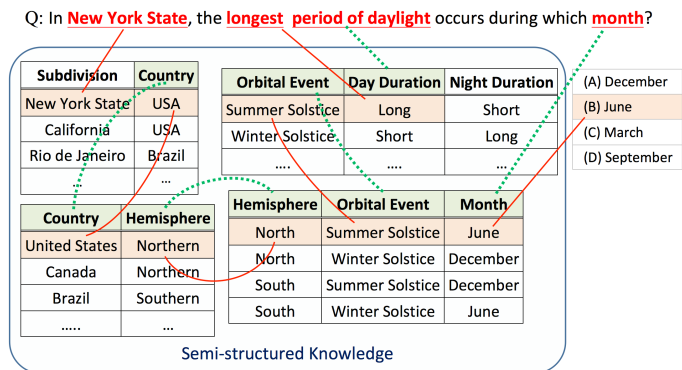
“better” support graphs = higher objective value

- Reward good behavior:
 - High lexical match links, nearby alignments, column header match, WH-terms (*“which of energy ...”*), etc.
- Penalize spurious overuse of frequently occurring terms

$$\max \sum_i c_i x_i$$

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

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



ILP Model: Some Details

Constraints:

- ~50 high-level constraints
- Scalability + consider only meaningful support graphs

- **Structural Constraints:**

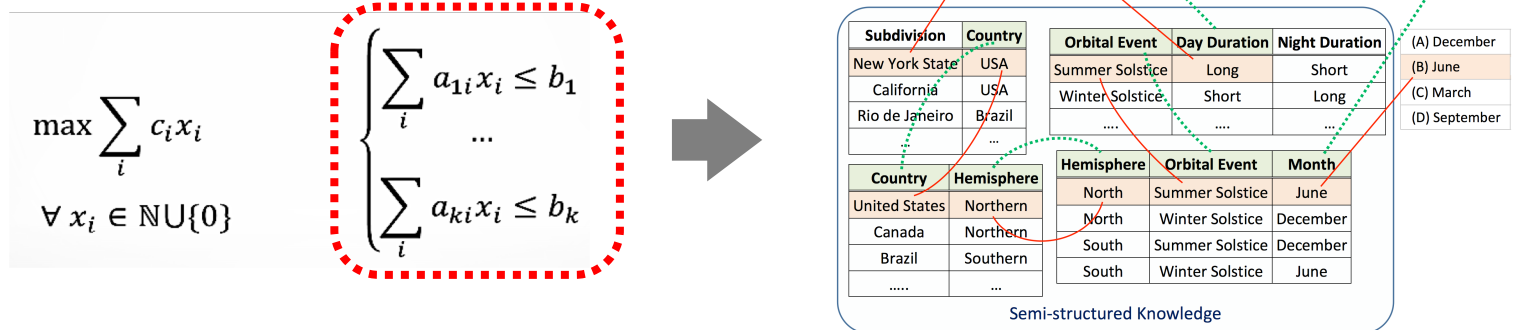
Many possible Proof Structures:

- Basic Lookup, Parallel Evidence, Evidence Chaining, Semantic Relation Matching

- **Semantic Constraints:**

Meaningful proof structures

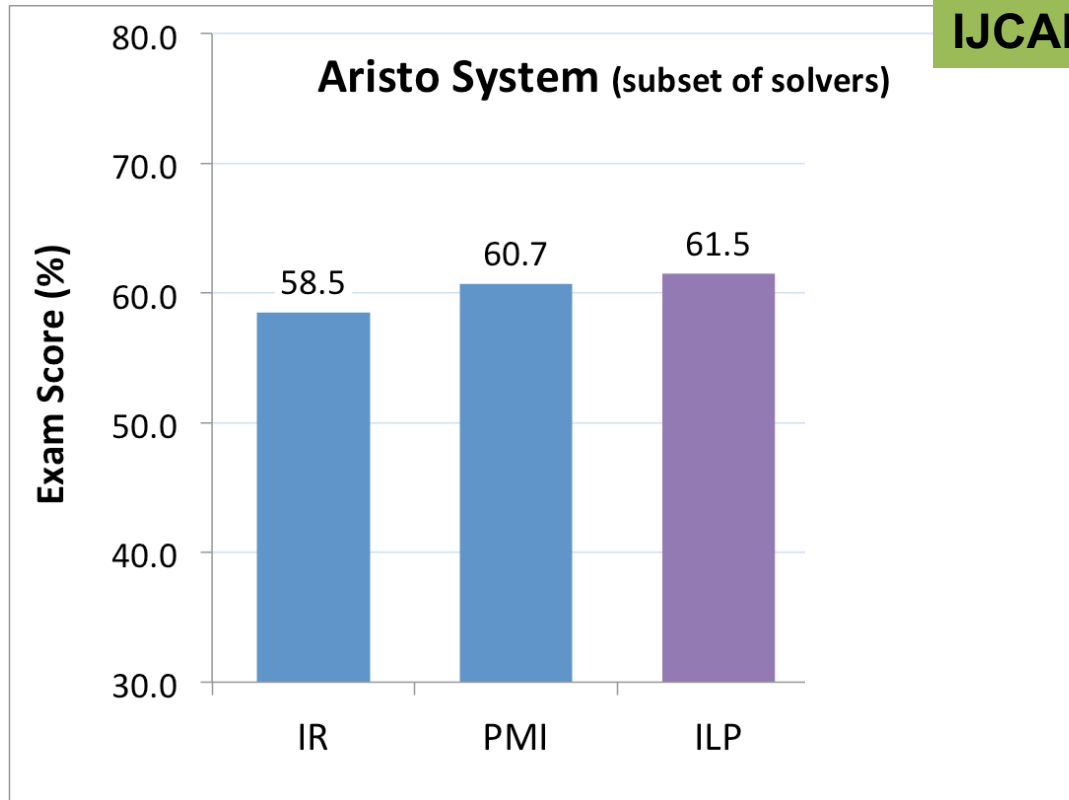
- Connectedness, question coverage, appropriate table use, etc.



Some Empirical Results



IJCAI-2016



More details in:
K et al, IJCAI-2016.

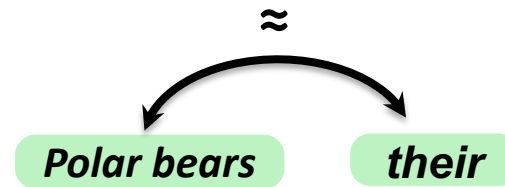
Ensemble performs 8-10% higher than IR baselines

Beyond Tables

- Issues with tabular representation:
 - Hard to extract; the schemas could be limiting
- Alternative representation:
 - Rich representation
 - Easy to automatically extract
- Idea: **Reason over (multiple) semantic abstractions of text**
 - Use off-the-shelf, pre-trained NLP modules for automatic extraction of representations
- Same ILP-based reasoning formalism as before

Mapping Text to Semantic Representations

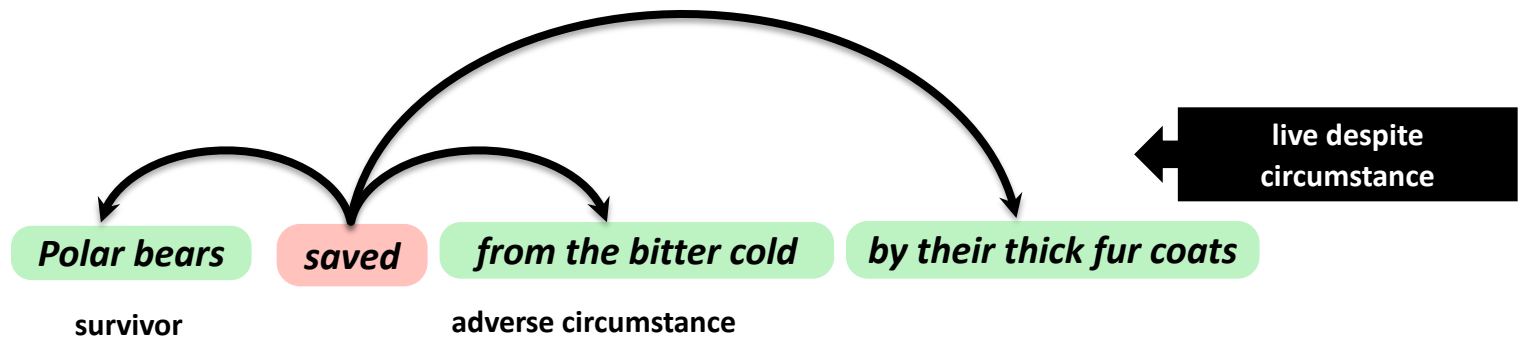
- NLP field has developed tools to extract and represent many interesting phenomena in language:
 - **Example 1: Co-reference**



*... **Polar bears**, saved from the bitter cold by **their** thick fur coats, are among the animals in danger of extinction ...*

Mapping Text to Semantic Representations

- NLP field has developed tools to extract and represent many interesting phenomena in language:
- **Example 3:** events described by verbs



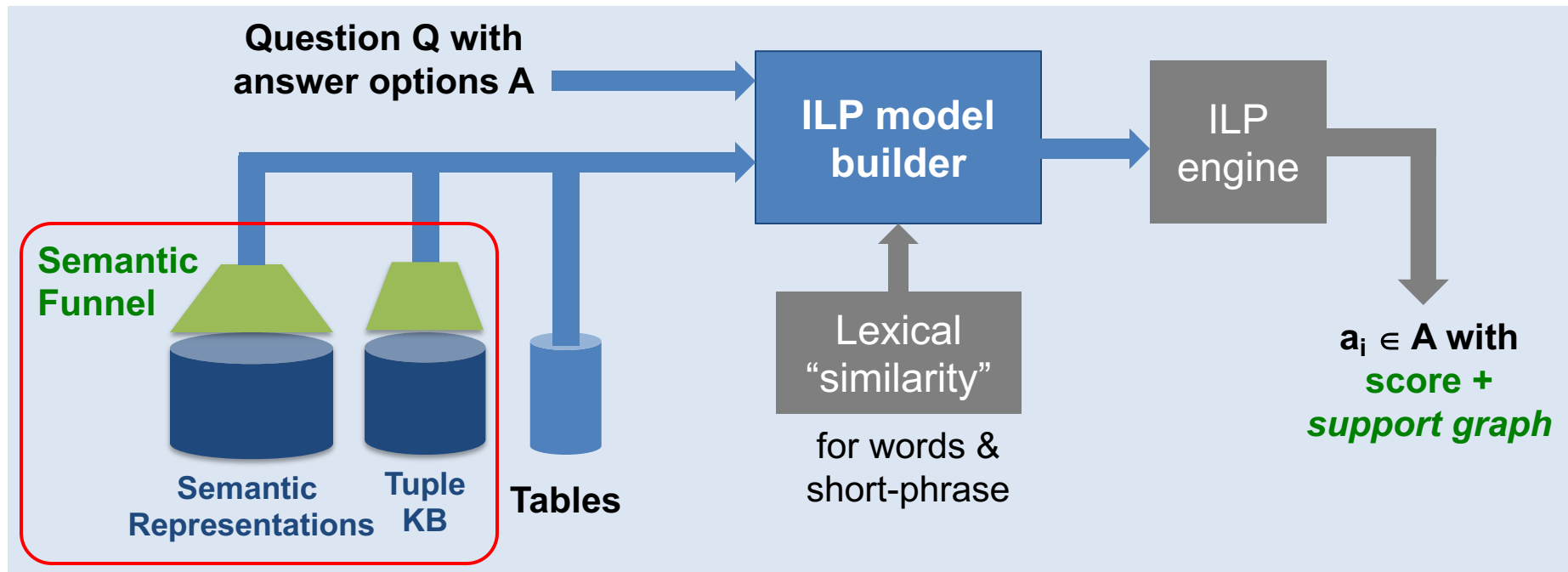
... Polar bears, saved from the bitter cold by their thick fur coats, are among the animals in danger of extinction ...

Verb Semantic Roles [Punyakanok et al. 2008]

Inference Solvers, Beyond Tables

- Extend model to consume various semantic representations.
 - *Trickier than it appears:* Can no longer rely on tables as a coherent collection of domain-relevant pieces of knowledge
 - Must revisit tuple selection, chaining, ...

Khushabi et al, AAAI-2018
Khot et al, ACL-2017

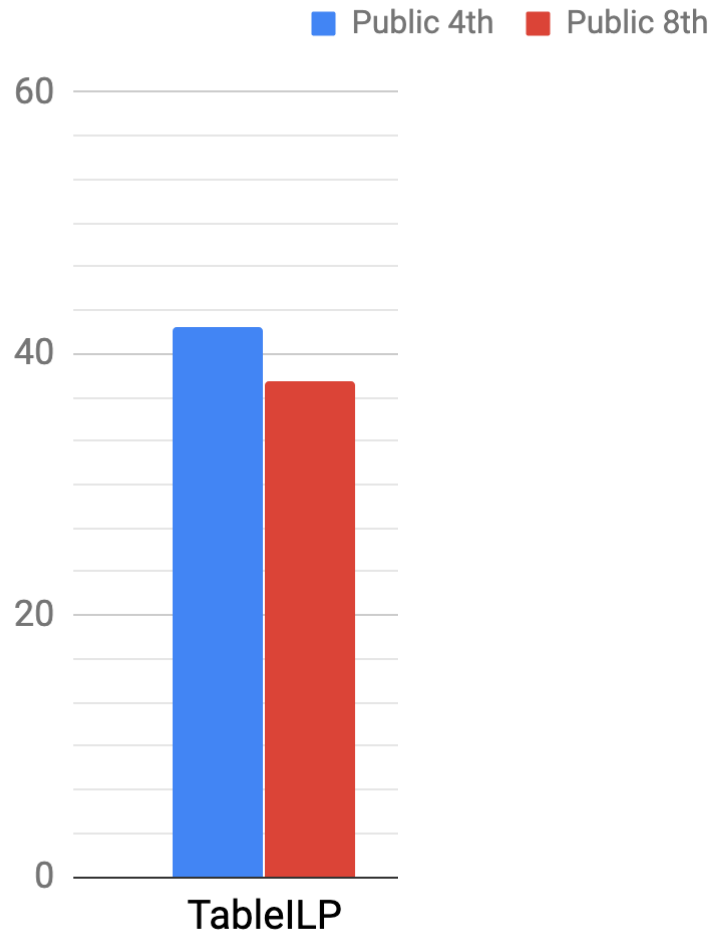


Some Empirical Results



AAAI-2018

ACL-2017



Top scoring systems
in their own time.

More details in
the related papers

Interim Summary

- The importance of the ability to chain information
- Structured, Multi-Step Reasoning:
 - Use science **knowledge in small, reusable, swappable pieces**
- **State-of-the-art results** on science benchmarks
- Benefits of the approach:
 - ✓ **Principled** approach
 - ✓ **Explainable** answers

Machine Reasoning [in Aristo]: Today and Future

Progression on NY Regents 8th Grade (NDMC)

THE UNIVERSITY OF THE STATE OF NEW YORK

GRADE 8

INTERMEDIATE-LEVEL
SCIENCE TEST

WRITTEN TEST

JUNE 6, 2011

Student Name _____

School Name _____

Print your name and the name of your school on the lines above.

The questions on this test measure your knowledge and understanding of science. The test has two parts. Both parts are contained in this test booklet.

Part I consists of 45 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 38 open-ended questions. Write your answers to these questions in the spaces provided in this test booklet.

You may use a calculator to answer the questions on the test if needed.

You will have two hours to answer the questions on this test.

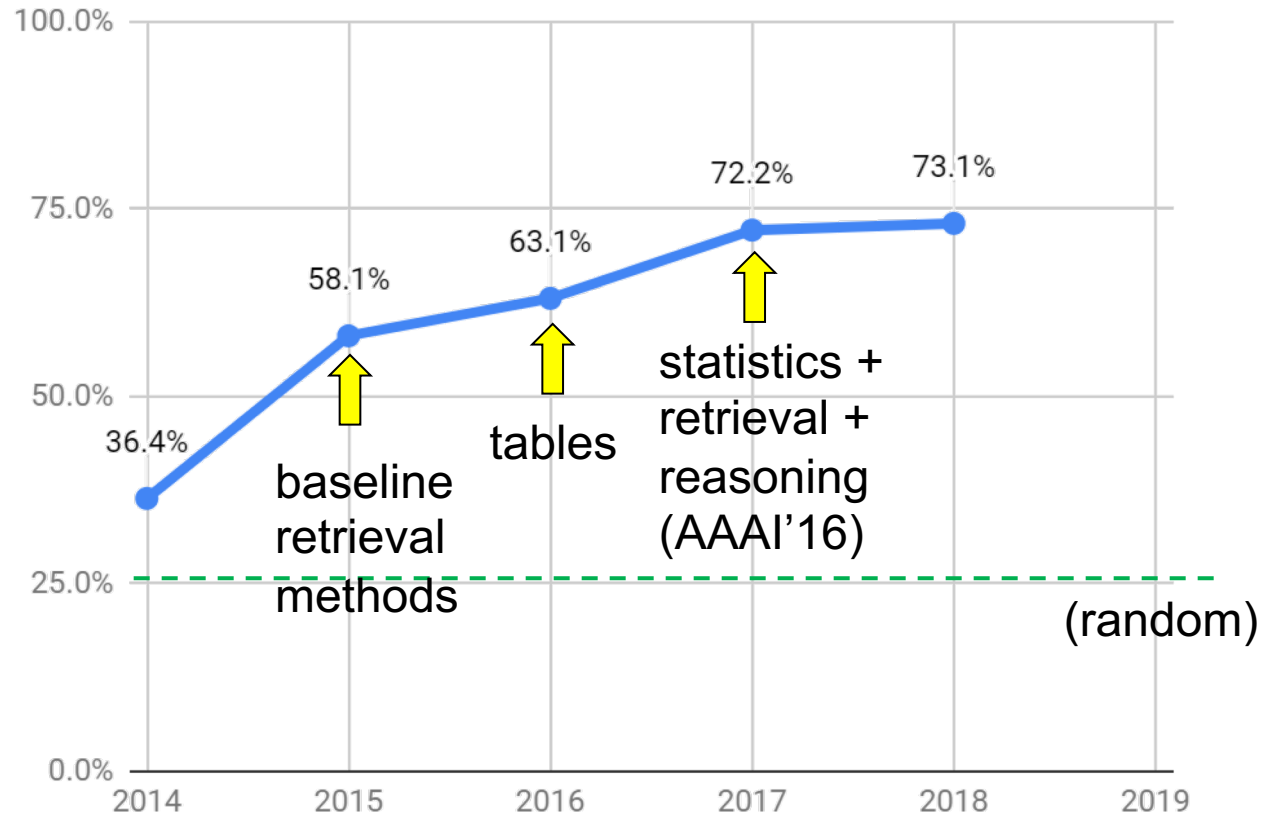
DO NOT TURN THIS PAGE UNTIL YOU ARE TOLD TO DO SO.

Copyright 2011

THE UNIVERSITY OF THE STATE OF NEW YORK

THE STATE EDUCATION DEPARTMENT

ALBANY, NEW YORK 12234



(hidden test set, questions as written, NDMC, 5 years/119 qns)

Progression on NY Regents 8th Grade (NDMC)

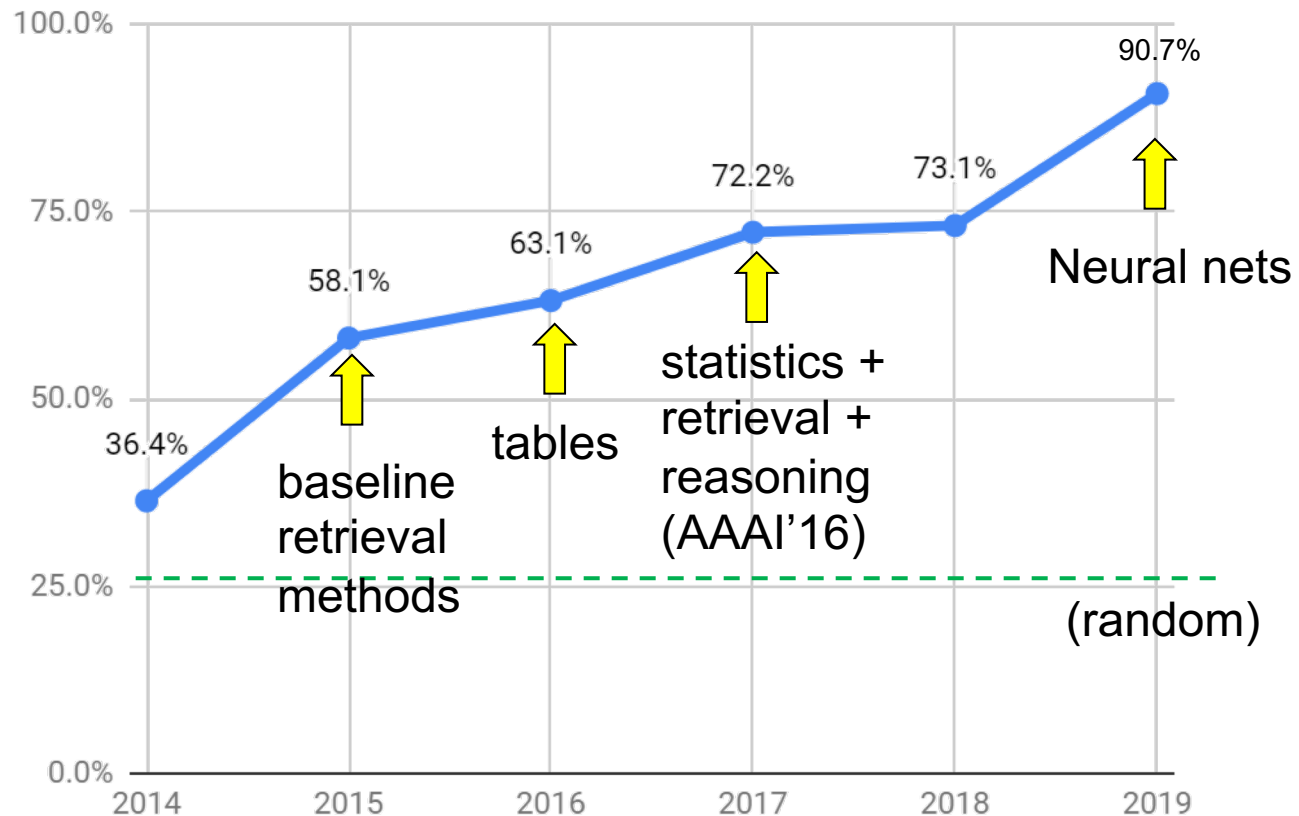
THE UNIVERSITY OF THE STATE OF NEW YORK
GRADE 8
INTERMEDIATE-LEVEL
SCIENCE TEST
WRITTEN TEST
JUNE 6, 2011

Student Name _____
School Name _____

Print your name and the name of your school on the lines above.
The questions on this test measure your knowledge and understanding of science.
The test has two parts. Both parts are contained in this test booklet.
Part I consists of 45 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 38 open-ended questions. Write your answers to these questions in the spaces provided in this test booklet.
You may use a calculator to answer the questions on the test if needed.
You will have two hours to answer the questions on this test.

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

Copyright 2011
THE UNIVERSITY OF THE STATE OF NEW YORK
THE STATE EDUCATION DEPARTMENT
ALBANY, NEW YORK 12234



Aristo aces 8th grade (non-diagram multiple choice) >90%

(hidden test set, questions as written, NDMC, 5 years/119 qns)

Aristo's Success



[SUBSCRIBE NOW](#) | [LOG IN](#)

A Breakthrough for A.I. Technology: Passing an 8th-Grade Science Test

By Cade Metz

Sept. 4, 2019

SAN FRANCISCO — Four scientists competed in a contest that could pass an eighth-grade science test for money on the line.

They all flunked. Even the best human team did better than 60 percent on the test, which tested science and logic skills that students learn in high school.

But on Wednesday, the Allen Institute for Artificial Intelligence, a prominent lab in Seattle, unveiled a new system that passed the test with room to spare. It correctly answered more than 90 percent of the questions on an eighth-grade science test and more than 80 percent on a 12th-grade exam.

Details in: [Clark et al, arXiv-2019](#)

From 'F' to 'A' on the N.Y. Regents Science Exams: An Overview of the Aristo Project*

**Peter Clark, Oren Etzioni, Daniel Khashabi, Tushar Khot, Bhavana Dalvi Mishra,
Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, Niket Tandon,
Sumithra Bhakthavatsalam, Dirk Groeneveld, Michal Guerquin, Michael Schmitz**

Allen Institute for Artificial Intelligence, Seattle, WA, U.S.A.

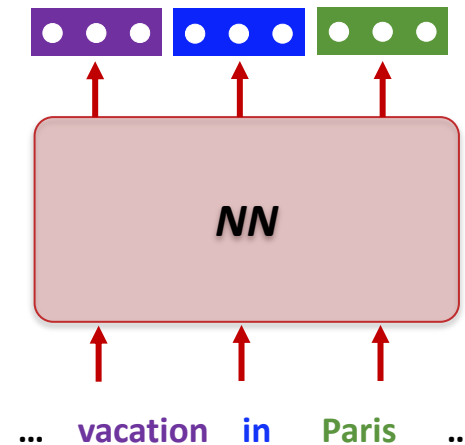
Abstract

AI has achieved remarkable mastery over games such as Chess, Go, and Poker, and even *Jeopardy!*, but the rich variety of standardized exams has remained a landmark challenge. Even in 2016, the best AI system achieved merely 59.3% on an 8th Grade science exam challenge (Schoenick et al., 2016).

What constraints are there on the interaction? What guidelines are provided to the judges? Second, recent Turing Test competitions have shown that, in certain formulations, the test itself is gameable; that is, people can be fooled by systems that simply retrieve sentences and make no claim of being intelligent (Aron, 2011; BBC, 2014). John Markoff of The New York Times wrote that the Turing Test is more

Neural Network solvers for natural language

- Big gains are achieved in the past 1-2 years
- Not just in Aristo: they have been pretty effective in a wide range of NLP problems



Many works:

[Peters et al. 18, Devlin et al.18; Radford et al.18, ...]

Neural Network solvers for natural language

- My view:
 - An engineering revolution
 - Concepts and ideas were out there (for decades)
- Enabled by:
 - ✓ **Big compute:**

Having access to many GPUs/TPUs is a must now
 - ✓ **Massive pre-training:**

Billions are documents are used to tune them
 - ✓ **Big supervision:**

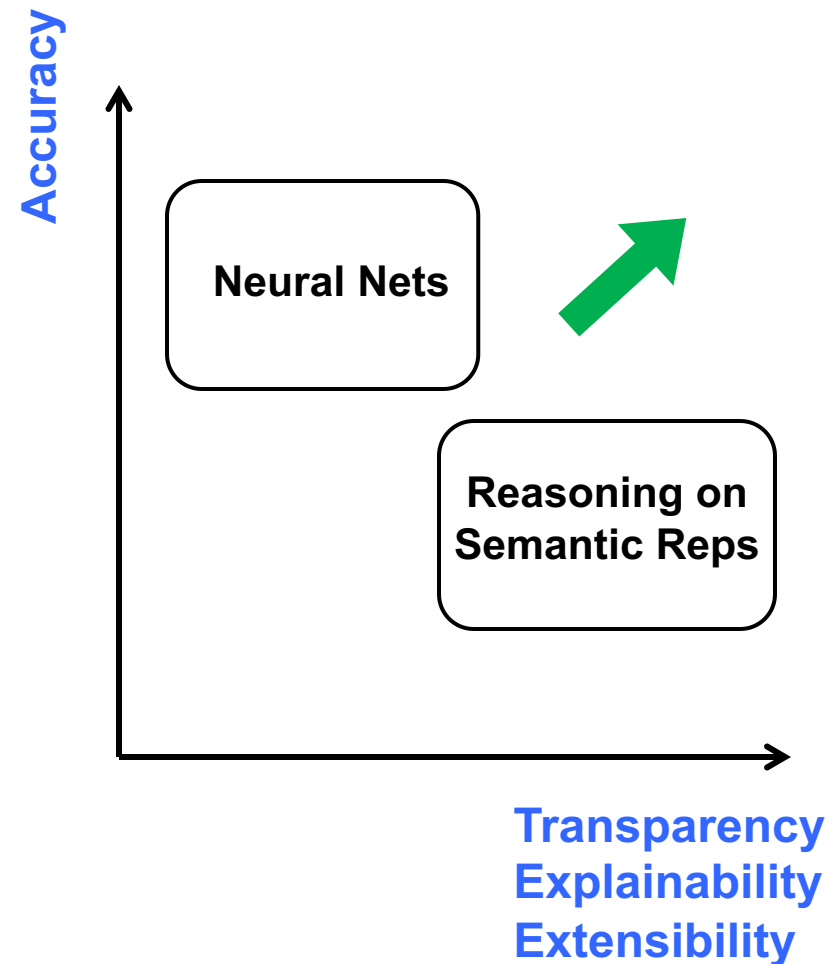
Availability of more annotated data

Neural Net resurgence



What's next? What is missing?

- Neural Nets have nearly “solved” the science challenges!
 - Do we understand why?
- In many cases, we are **not** able to **explain** their decisions.
 - As a result, we can't assess the extent of their **reliability**.
- Additionally, **extending** them is not trivial (largely an open problem)



Aristo: Probing the predictions

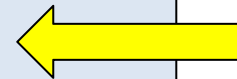
City administrators can encourage energy conservation by

- (1) lowering parking fees
- (2) building larger parking lots
- (3) decreasing the cost of gasoline
- (4) lowering the cost of bus and subway fares



City administrators can encourage energy conservation by

- (1) lowering parking fees
- (2) building larger parking lots
- (3) ~~decreasing~~ the cost of gasoline
- (4) ~~lowering~~ the cost of bus and subway fares



increasing
increasing

So maybe it does actually work? 🤔

Aristo: Probing the predictions

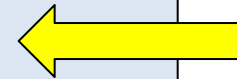
City administrators can encourage energy conservation by

- (1) lowering parking fees
- (2) building larger parking lots
- (3) decreasing the cost of gasoline
- (4) lowering the cost of bus and subway fares



City administrators can encourage energy conservation by

- (1) lowering parking fees
- (2) building larger parking lots
- (3) decreasing the cost of gasoline
- (4) ~~lowering the cost of bus and subway fares~~
increasing availability



Hmm 🤔 ... but how should we fix this??

Summary

- The Aristo experiment:
 - Working hypothesis: Real-world language and reasoning capabilities can be assessed via “well-designed” QA tasks.
- First systems for science QA effective at multi-hop inference
 - Can operate with semi-formal knowledge bases
 - Can fruitfully exploit noisier & less structured representations
 - Complementary to other methods
- Deep learning is playing an important role, but is not everything (yet)
- Interacting intelligently with humans remains challenging!
 - Requires rich linguistic understanding, common sense, domain knowledge, situational awareness, conversational memory ...

Questions?

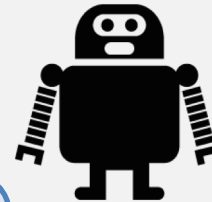


Hiring talented interns, young investigators, researchers, engineers, and more!



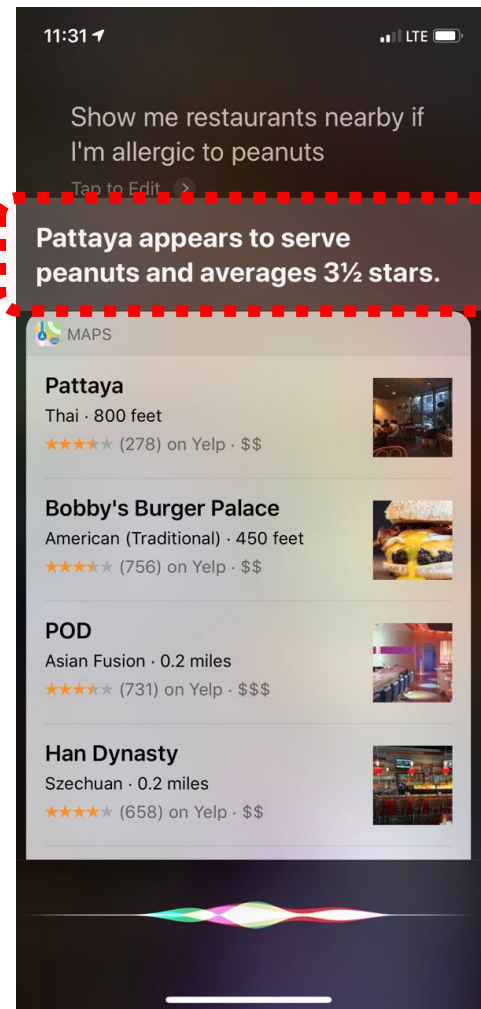
*Show me some
restaurants
nearby.*

*I don't like
crowds.*



*Here are some
options I found
nearby:*

....



What needs to happen?

1. Define representation of functions
2. Manually author and automatically learn extraction rules
3. Attack challenges

- Linguistic variability

- “synthesize protein”, “protein synthesis”, “synthesizes large numbers of proteins”, “helps to synthesize proteins”, ...



- Multiword expressions

- “cell membrane” = “membrane”, but “amino acid” ≠ “acid”



- Ambiguous patterns

- “cell membrane”, “cell division”, “cell biologist”



- Non-functional actions

- exist, differ, attempt, do, remain, appear, occur, play, show, ...



- Peripheral activities

- nucleus + divide, ribosome + move



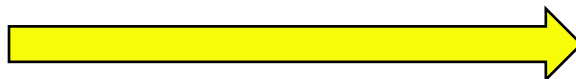
Aristo in Context

Watson

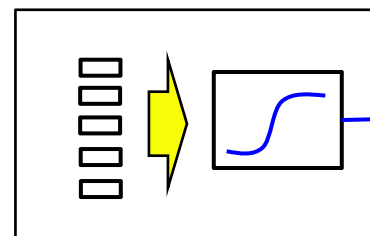
Text



candidate answer extraction



candidate
answer
phrases answer scoring +
evidence
combination



ranked answers

95%
34%
:

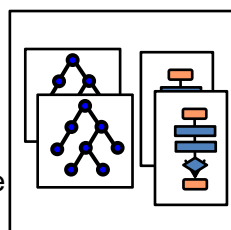
Aristo

Text

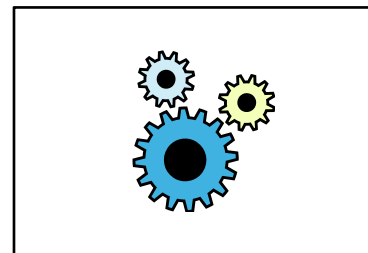


Knowledge Base
(library of ontologies and
domain models)

Natural Language
Processing



Tractable, probabilistic
Inference



answer explanation

□, because

Cyc

Knowledge
Engineers



Knowledge Base
(monolithic,
fixed ontology)

```
(thereExists ?VAR0
  (and
    (chemicalReactants-TypeType
      CalvinCycle ?VAR1)
    (isa ?VAR1 ChemicalCompound
      TypeByChemicalSpecies)
    (different CarbonDioxide ?VAR1)
    (chemicalProducts-TypeType
      CalvinCycle ?VAR0)
    (genls ?VAR0 CarbohydrateStuff))
    (isa genls TransitiveBinaryPredicate)
    (genls Sugar CarbohydrateStuff)
    (isa (ChemicalSubstanceFn ...
```

Heavy-duty deductive inference



answer explanation

□, because

Example

QUESTION

Water freezing is an example of a (A) liquid turning into a solid (B) solid turning into a liquid (C) gas turning into a solid

NEW: object(freeze,water) -> d-change(water ,-, liquid, solid)?

object prop from to
↑
explicit representation of discrete change

TEXT

Freezing involves changing water from its liquid state to its solid state by the removal of heat.

freeze -> d-change(water, -, liquid state, solid state).
object property from to

Example

QUESTION

Water freezing is an example of a (A) liquid turning into a solid (B) solid turning into a liquid (C) gas turning into a solid

NEW: `object(freeze, water) -> d-change(water, -, liquid, solid)?`

explicit representation of discrete change

TEXT

Freezing involves changing water from its liquid state to its solid state by the removal of heat.

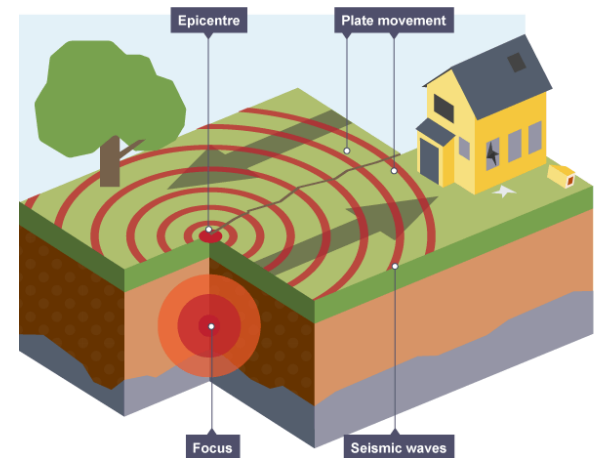
`freeze -> d-change(water, -, liquid state, solid state).`



REASONING

What do earthquakes tell scientists about the history of the planet?

- (A) Earth's climate is constantly changing.
- (B) The continents of Earth are continually moving.
- (C) Dinosaurs became extinct about 65 million years ago.
- (D) The oceans are much deeper today than millions of years ago.





ARISTO ANSWERED:

Question: What do earthquakes tell scientists about the history of the planet?

Hide

Answer: (C) Dinosaurs became extinct about 65 million years ago.

Confidence: 45.42%

as computed from these reasoners:

Information Retrieval: 87.59%

Table Reasoning: 6.68%

Topic Matching: 65.20%

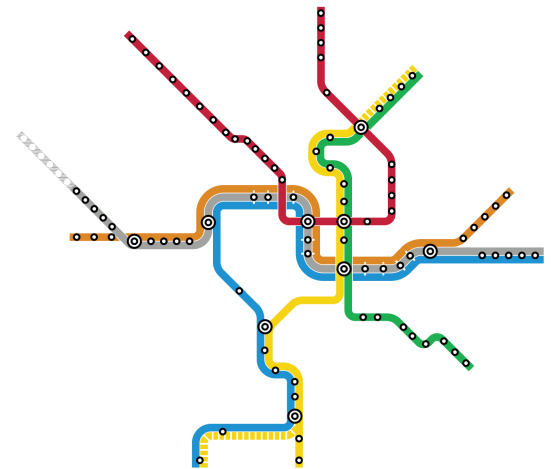
Tuple Reasoning: 4.16%

Justification sentence (from Information Retrieval): The paleontological records of the history of life on this planet show that the giant dinosaurs and many other animal and plant groups became extinct about 65 million years ago.

COMMON SENSE

City administrators can encourage energy conservation by

- (A) lowering parking fees
- (B) building larger parking lots
- (C) decreasing the cost of gasoline
- (D) lowering the cost of bus and subway fares



ARISTO'S BEST GUESS:

Question: City administrators can encourage energy conservation by

Hide

Aristo is not sure about this one...

Best guess: (C) decreasing the cost of gasoline

Confidence: 35.55%

as computed from these reasoners:

Information Retrieval: 14.95%

Topic Matching: 70.15%

Tuple Reasoning: 69.43%

Justification sentence (from Information Retrieval): 1970s Programs and educational materials are created to encourage gasoline and energy conservation.

Reasoning to an Answer: Qualitative Relations

How are the particles in a block of iron affected when the block is melted?

- (A) The particles gain mass.
- (B) The particles contain less energy.
- (C) The particles move more rapidly.
- (D) The particles increase in volume.



Reasoning to an Answer: Qualitative Relations

How are the particles in a block of iron affected when the block is melted?

- (A) The particles gain mass.
- (B) The particles contain less energy.
- (C) The particles move more rapidly.
- (D) The particles increase in volume.



RETRIEVED KNOWLEDGE

As the heat of a particle increases, the particles move faster.



Retrieved knowledge



Reasoning to an Answer: Qualitative Relations

How are the **particles** in a block of iron affected when the block is **melted**?

- (A) The particles gain mass.
- (B) The particles contain less energy.
- (C) The particles move **more rapidly**.
- (D) The particles increase in volume.



RETRIEVED KNOWLEDGE

As the **heat** of a **particle** increases, the particles **move faster**.

IDENTIFIED SPANS FOR QUALITATIVE PROPERTY:

particles (0.77)

particle (0.65)

heat (0.62)

IDENTIFIED SPANS FOR QUALITATIVE DIRECTION:

more rapidly (0.93)

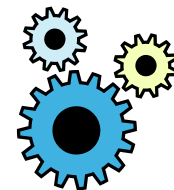
faster (0.93)

increases (0.93)

Retrieved knowledge



Application to question



Example Question: Simple Lookup

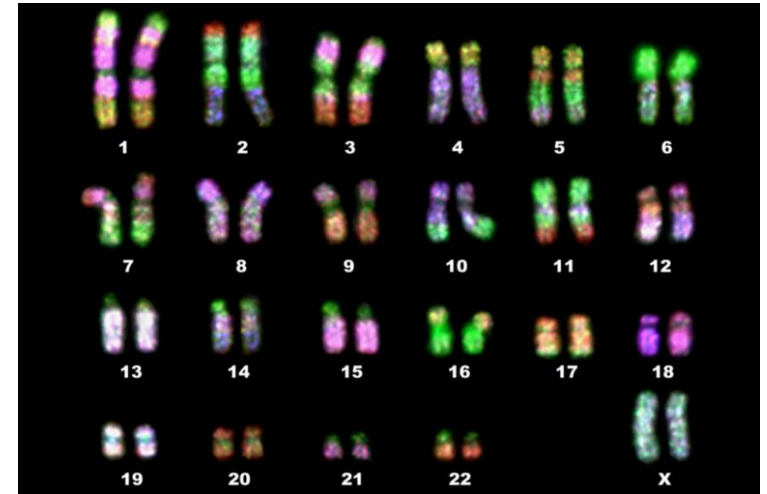
How many chromosomes does the human body cell contain?

(A) 23

(B) 32

(C) 46

(D) 64



Example Question: Reasoning by Chaining

Which features are the best evidence that **glaciers** once covered an area?

- (A) wide riverbeds
- (B) U shaped **valleys**
- (C) groundwater springs
- (D) underground caves



Aristo KB facts:

(**glacier**, causes, glacial erosion)

&

(glacial erosion, causes, **valley**)

→ (**glacier**, causes, **valley**)

Handling Lexical Variability



AAAI-2018

ACL-2017

Knowledge is expressed in a **variety of linguistic forms**

- Simple textual variation confuses even the best solvers
- No single knowledge representation (e.g., Open IE tuples) suffices
- **Idea: Reason over (multiple) semantic abstractions of text**
 - Use off-the-shelf, pre-trained NLP modules
 - Multiple views for a more complete semantic understanding
 - SRL frames (verbs, prepositions, comma), dependency parse, coreference sets, lexical similarity links, raw text sequence
- Unified representation as a family of **Semantic Graphs**
 - PredArg graphs, trees, clusters, sequences
 - Connected via textual similarity links
- Same ILP-based reasoning formalism as before