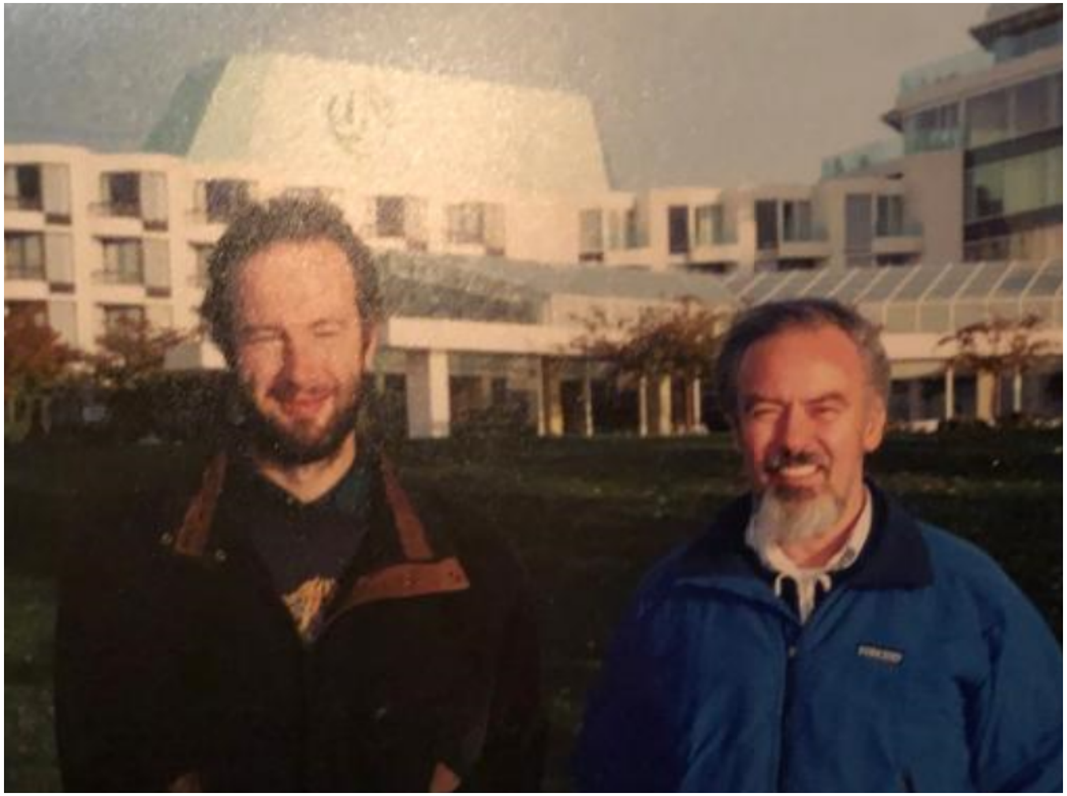
Project Aristo: Towards Machines that Capture and Reason with Science Knowledge

Peter Clark

November 2019

The History of KCap

- KCap 2001-19



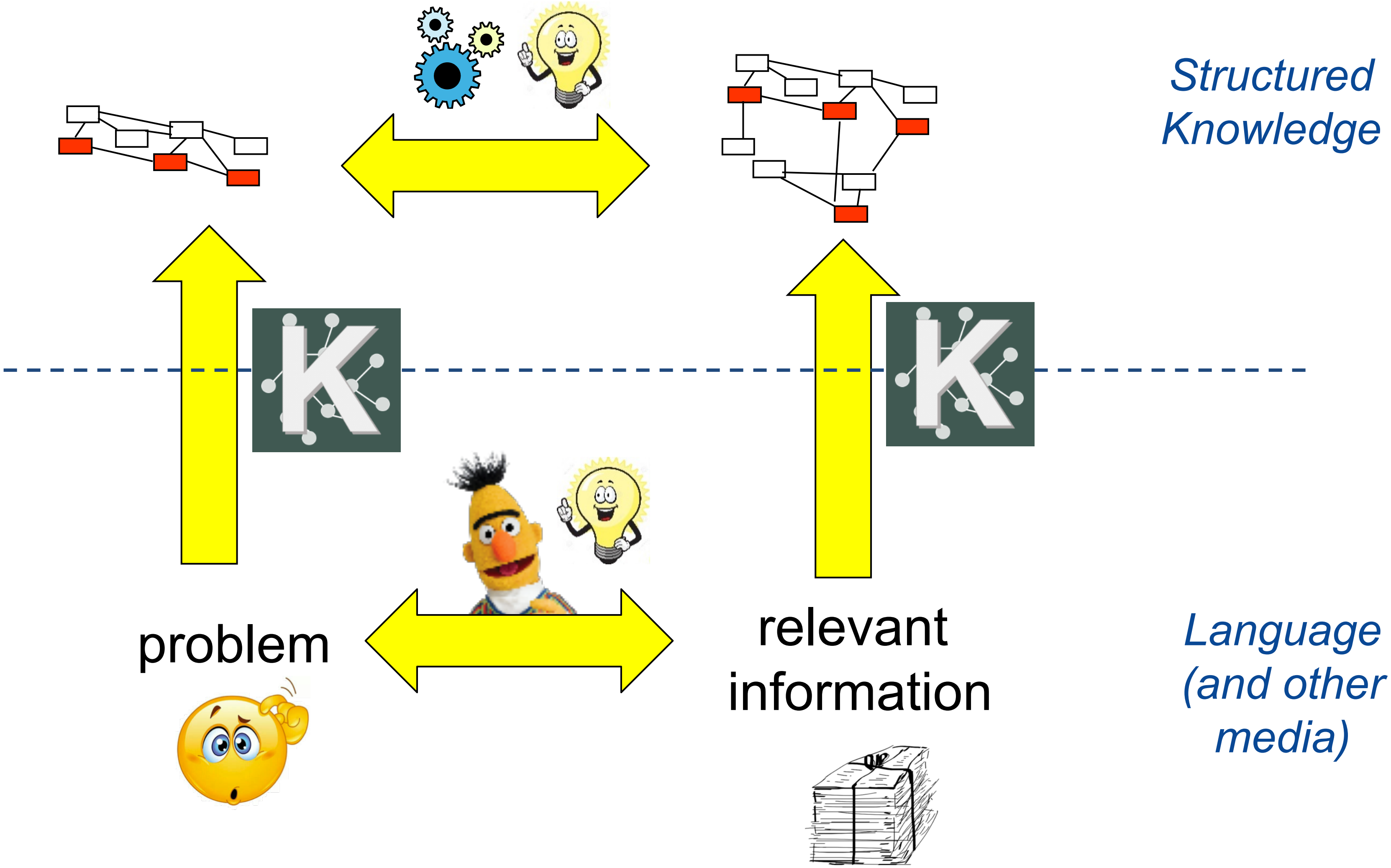
The History of KCap

- KCap 2001-19

- Banff Knowledge Acquisition Workshops: 1986-1999

How do we get knowledge into the machine in a usable form?

A Question for the field of Knowledge Capture



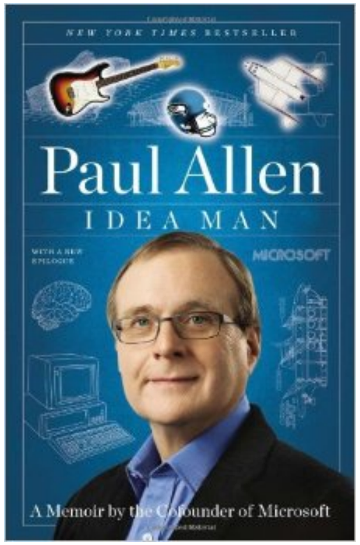
problem

relevant information

Structured Knowledge

Language (and other media)

Science Questions: A Grand Challenge…


NEW YORK TIMES BESTSELLER
Paul Allen
IDEA MAN
MICROSOFT
A Memoir by the Cofounder of Microsoft

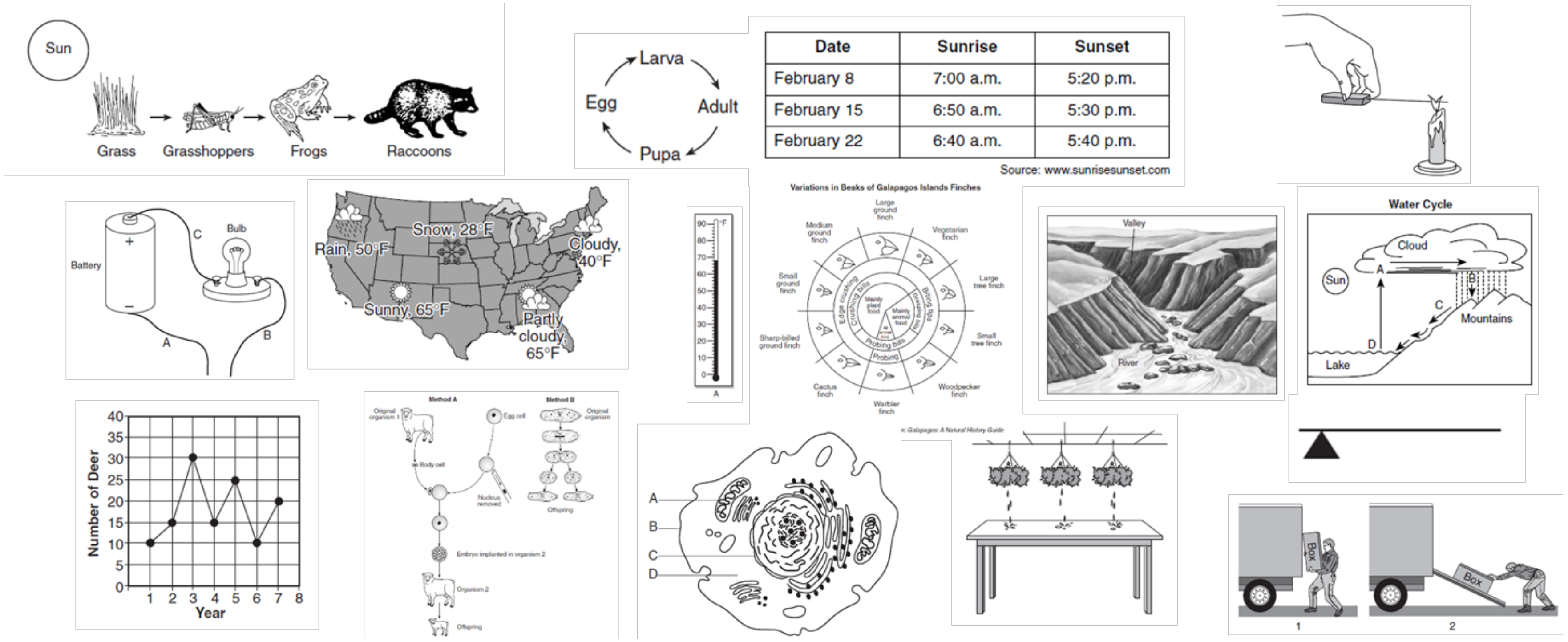
Over the last decade, I began to think about a ***"Digital Aristotle",*** *an easy-to-use, all-encompassing knowledge storehouse....to advance the field of AI.*

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.

Question Categories Not Covered

- Diagrams

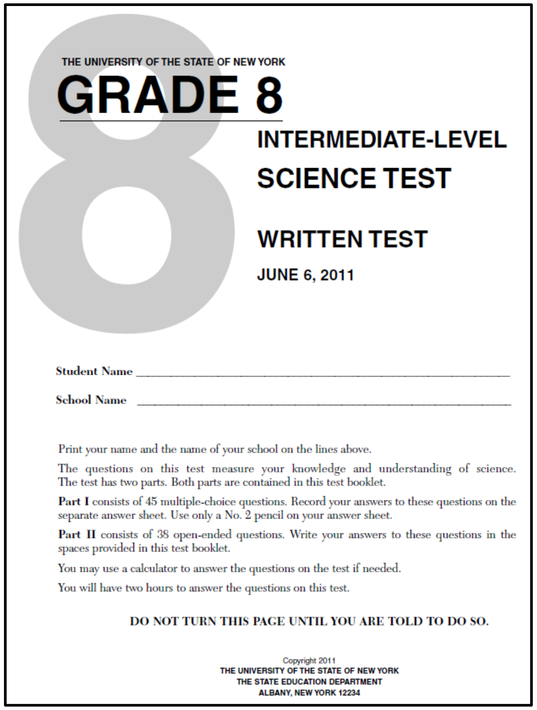


Date	Sunrise	Sunset
February 8	7:00 a.m.	5:20 p.m.
February 15	6:50 a.m.	5:30 p.m.
February 22	6:40 a.m.	5:40 p.m.

Source: www.sunrisesunset.com

- Direct Answer Questions

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.

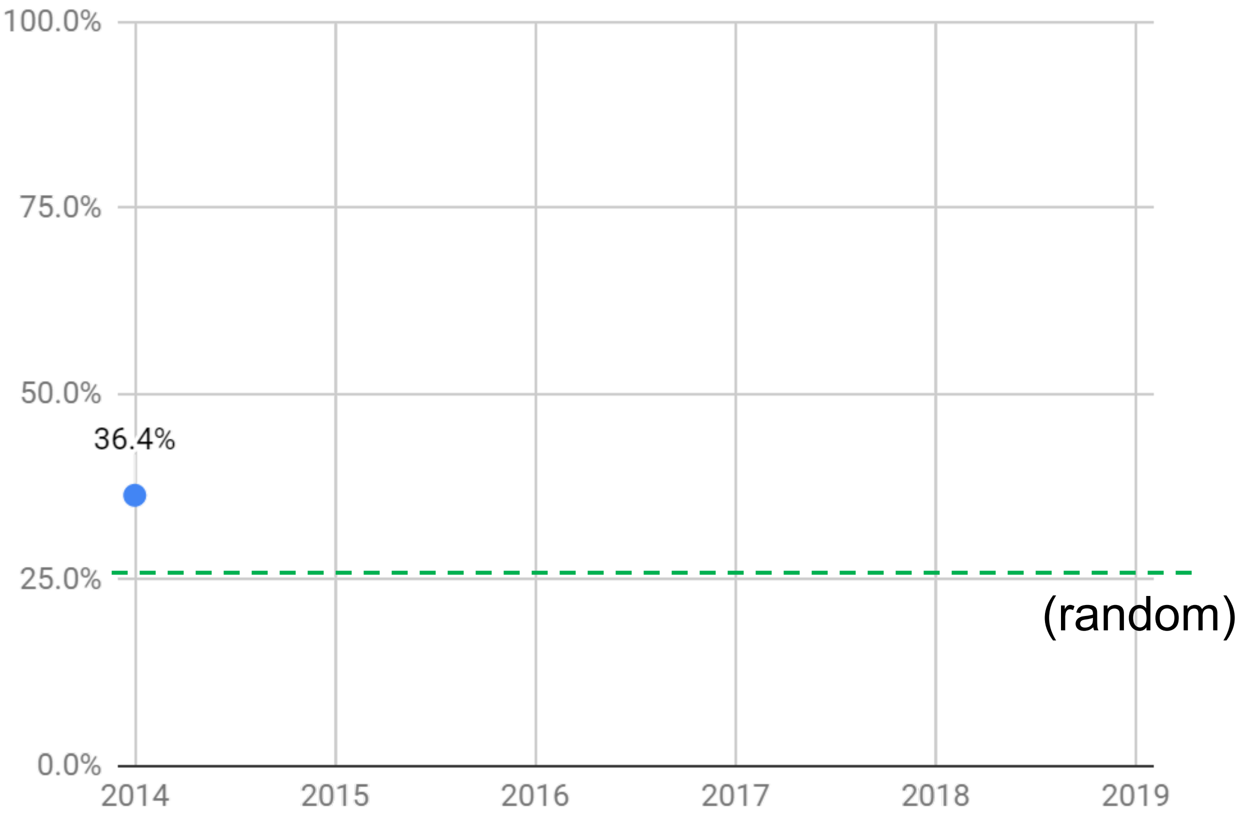
Part I consists of 45 multiple-choice questions. Record your answers to these questions on the separate answer sheet. Use only a No. 2 pencil on your answer sheet.

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.

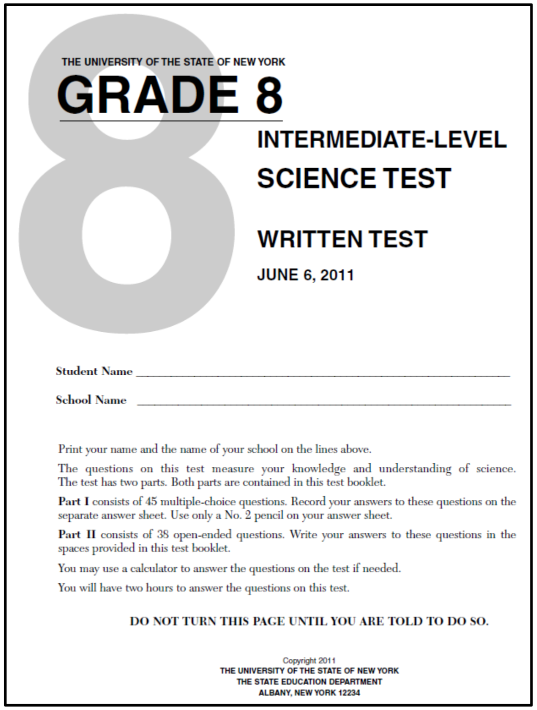
DO NOT TURN THIS PAGE UNTIL YOU ARE TOLD TO DO SO.



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

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

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.

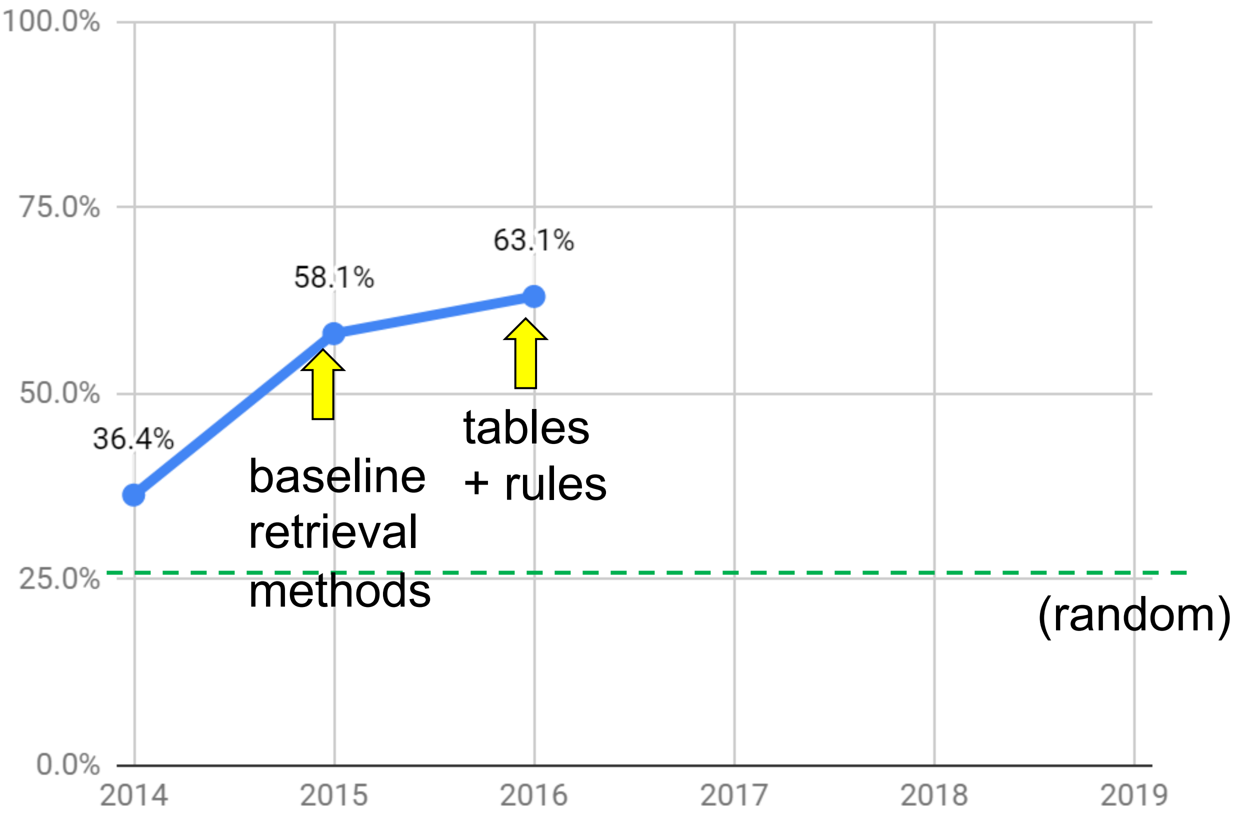
Part II consists of 38 open-ended questions. Write your answers to these questions in the spaces provided in this test booklet.

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

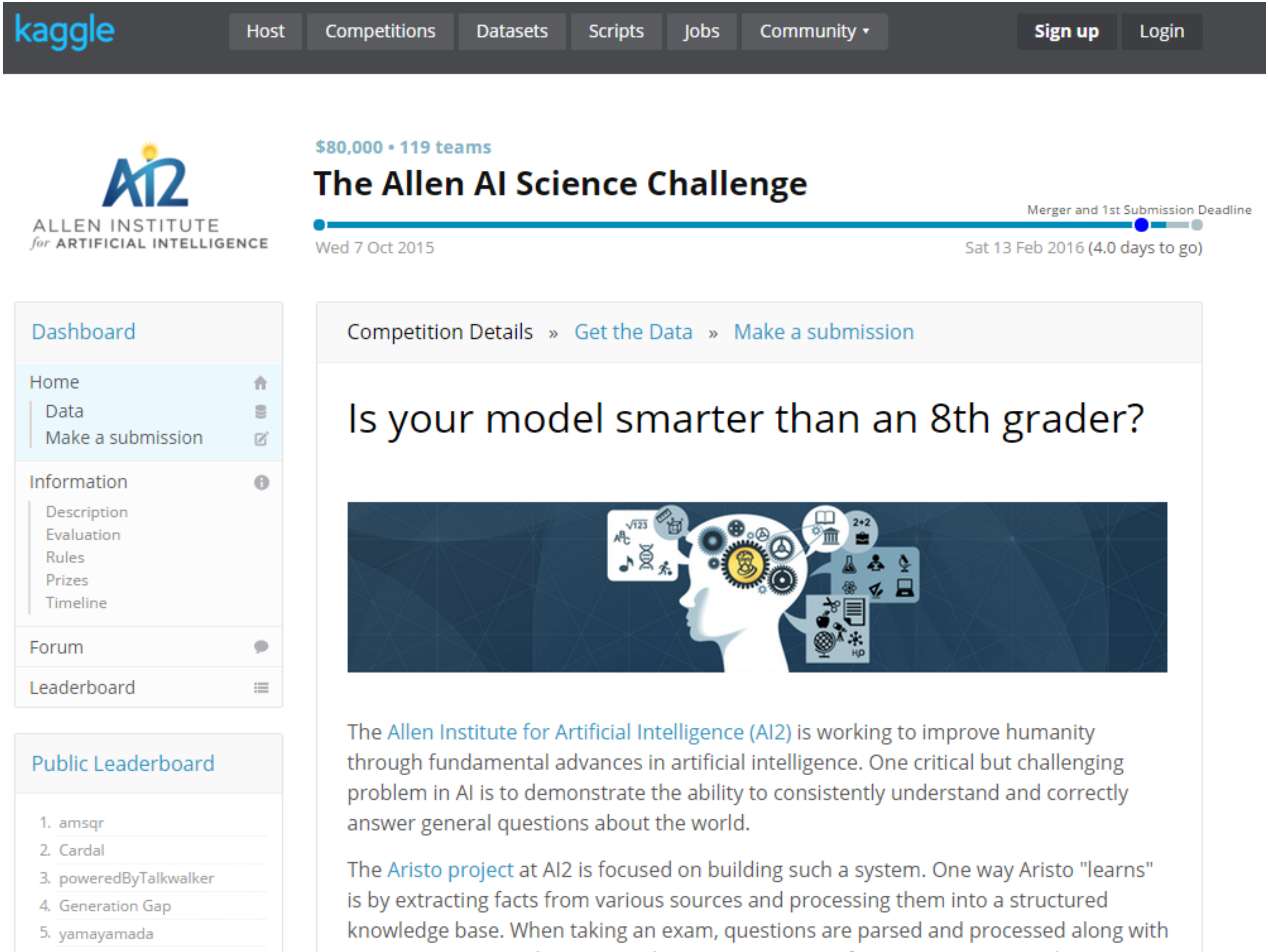


baseline retrieval methods

tables + rules

(random)

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



$80,000 • 119 teams

The Allen AI Science Challenge

Merger and 1st Submission Deadline

Wed 7 Oct 2015

Sat 13 Feb 2016 (4.0 days to go)

Public Leaderboard

1. amsqr
2. Cardal
3. poweredByTalkwalker
4. Generation Gap
5. yamayamada

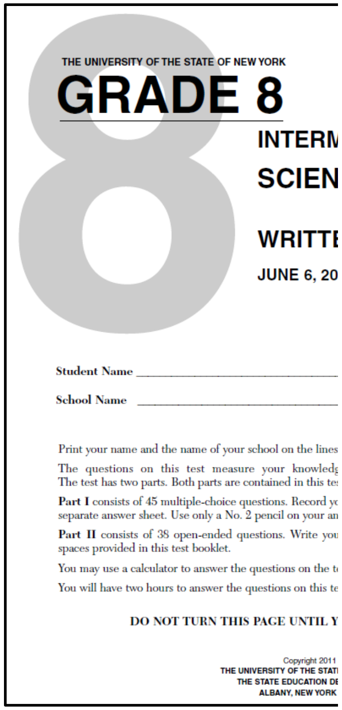
Competition Details » Get the Data » Make a submission

Is your model smarter than an 8th grader?

The Allen Institute for Artificial Intelligence (AI2) is working to improve humanity through fundamental advances in artificial intelligence. One critical but challenging problem in AI is to demonstrate the ability to consistently understand and correctly answer general questions about the world.

The Aristo project at AI2 is focused on building such a system. One way Aristo "learns" is by extracting facts from various sources and processing them into a structured knowledge base. When taking an exam, questions are parsed and processed along with

Progression on NY Regents 8th Grade (NDMC)



THE UNIVERSITY OF THE STATE OF NEW YORK

GRADE 8

INTERM

SCIEN

WRITTE

JUNE 6, 20

Student Name ______________________

School Name ______________________

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

The questions on this test measure your knowledg
The test has two parts. Both parts are contained in this te

Part I consists of 45 multiple-choice questions. Record yo
separate answer sheet. Use only a No. 2 pencil on your an

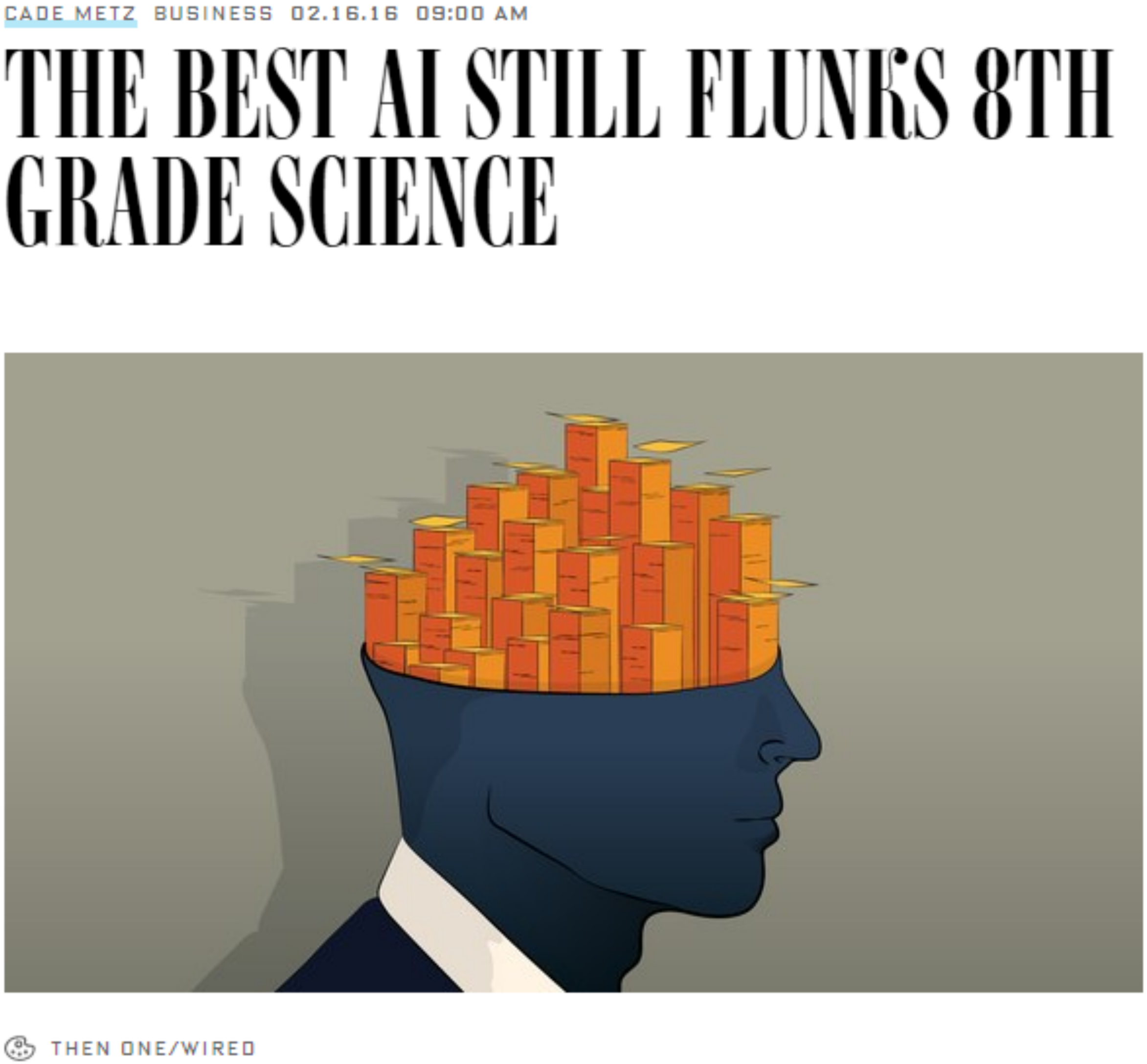
Part II consists of 38 open-ended questions. Write you
spaces provided in this test booklet.

You may use a calculator to answer the questions on the t

You will have two hours to answer the questions on this te

DO NOT TURN THIS PAGE UNTIL Y

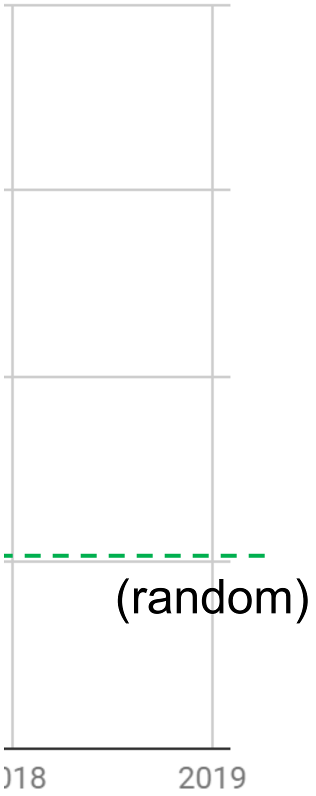
Copyright 2011
THE UNIVERSITY OF THE STAT
THE STATE EDUCATION DE
ALBANY, NEW YORK



CADE METZ BUSINESS 02.16.16 09:00 AM

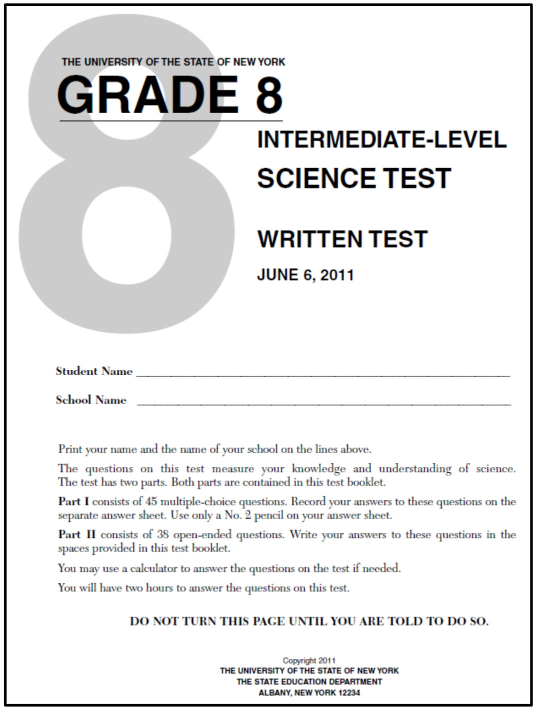
THE BEST AI STILL FLUNKS 8TH GRADE SCIENCE

THEN ONE/WIRED



(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.

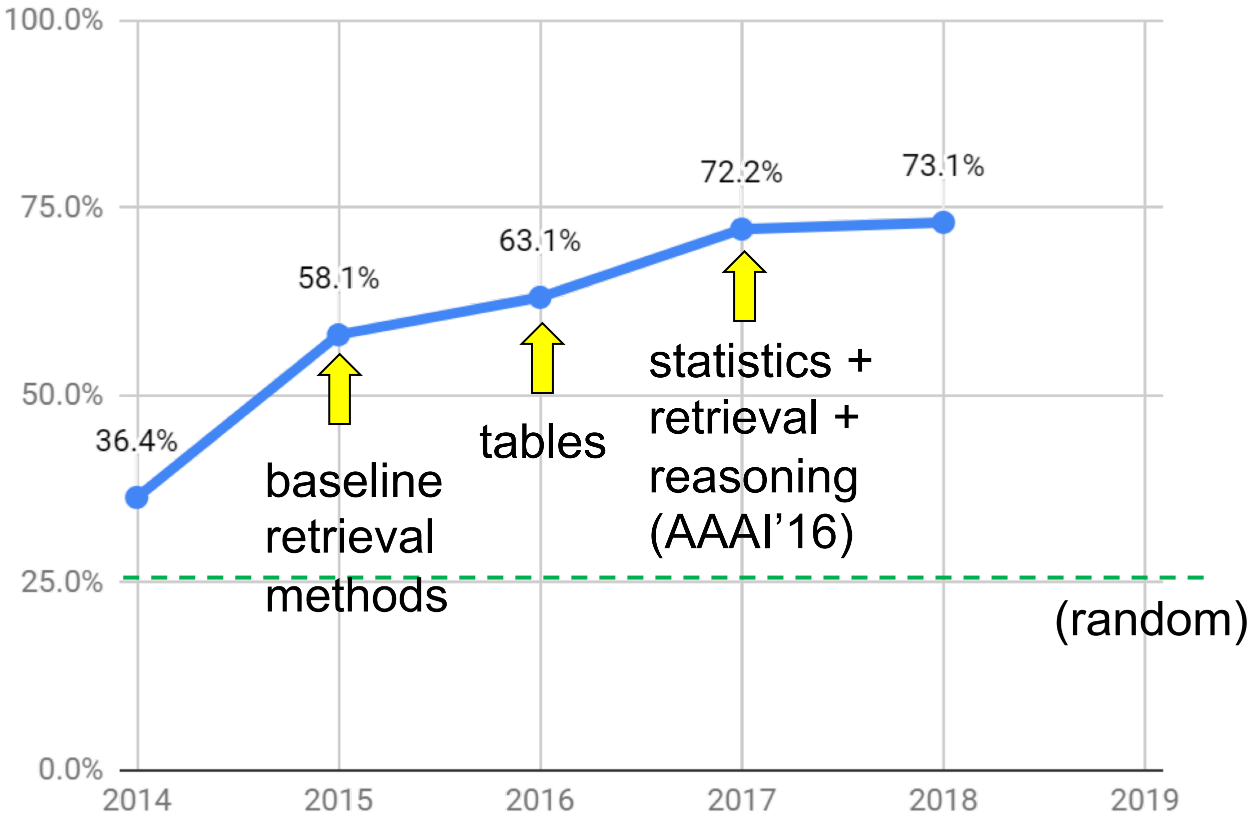
Part II consists of 38 open-ended questions. Write your answers to these questions in the spaces provided in this test booklet.

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



baseline retrieval methods

tables

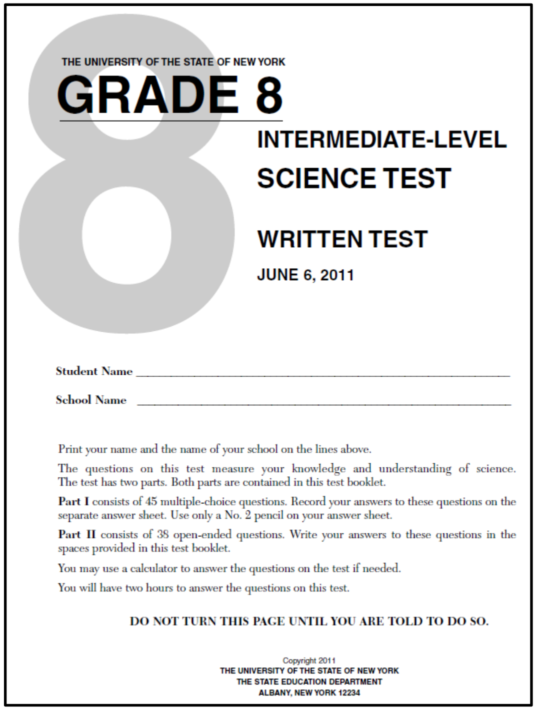
statistics + retrieval + reasoning (AAAI'16)

(random)

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

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

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.

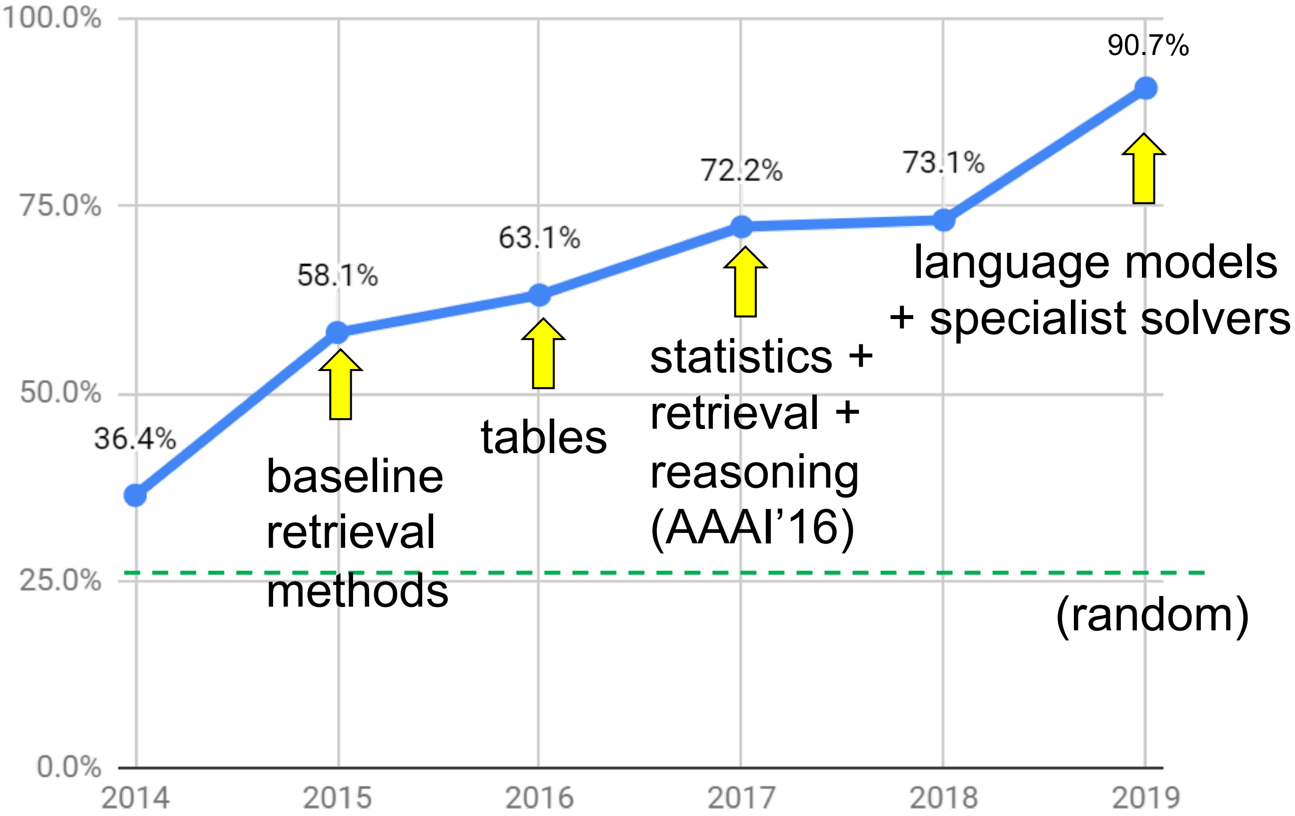
Part II consists of 38 open-ended questions. Write your answers to these questions in the spaces provided in this test booklet.

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



baseline retrieval methods

tables

statistics + retrieval + reasoning (AAAI'16)

language models + specialist solvers

(random)

Separate test on 3 latest exams (2017-2019): 93.3%

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

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

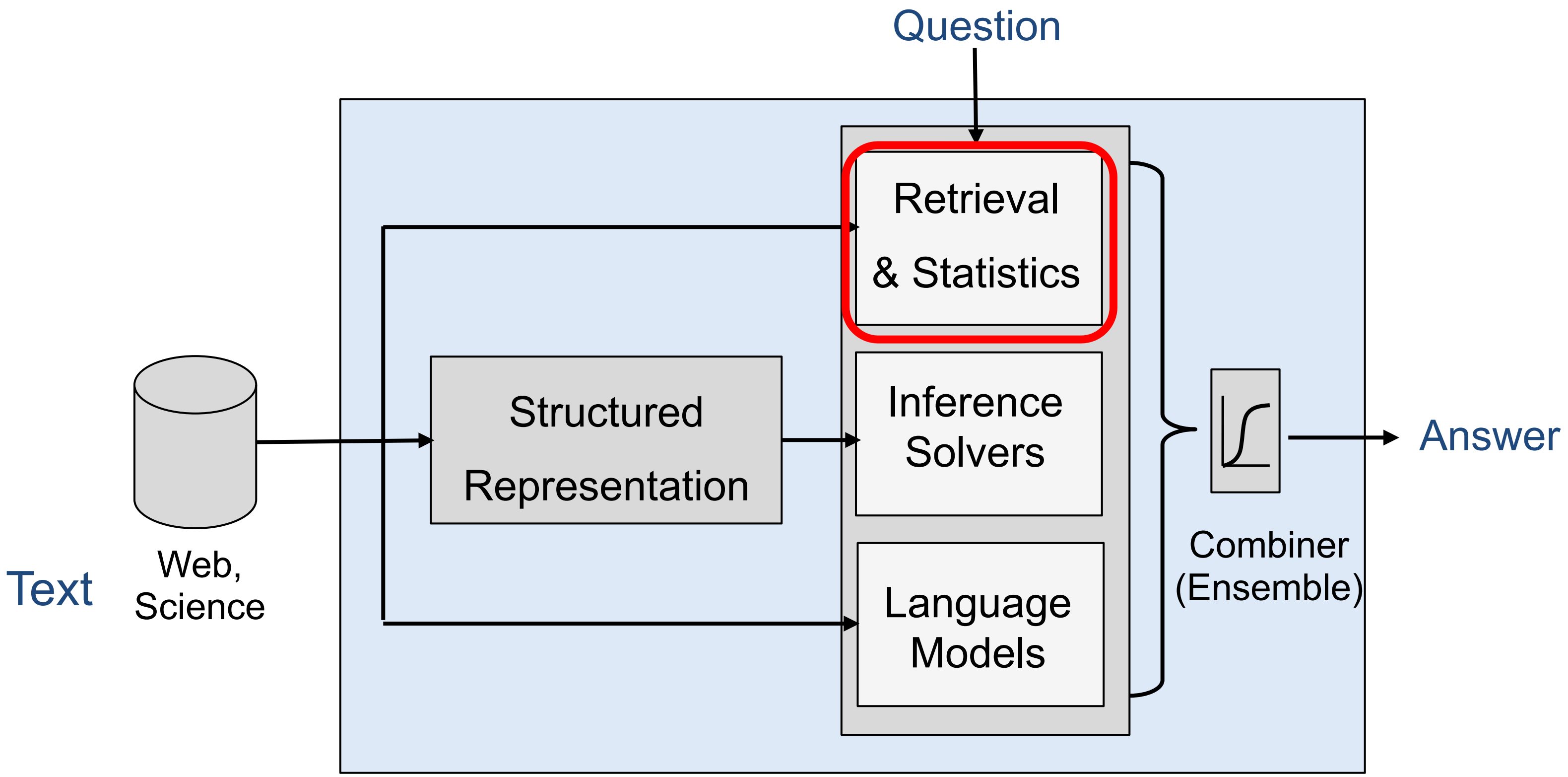
Outline

- Introduction
- How does Aristo work? ⇦
- What is going on behind the high scores on the exams?
- Where does Aristo fail?
- What are steps forward?

Aristo: an over-simplified overview

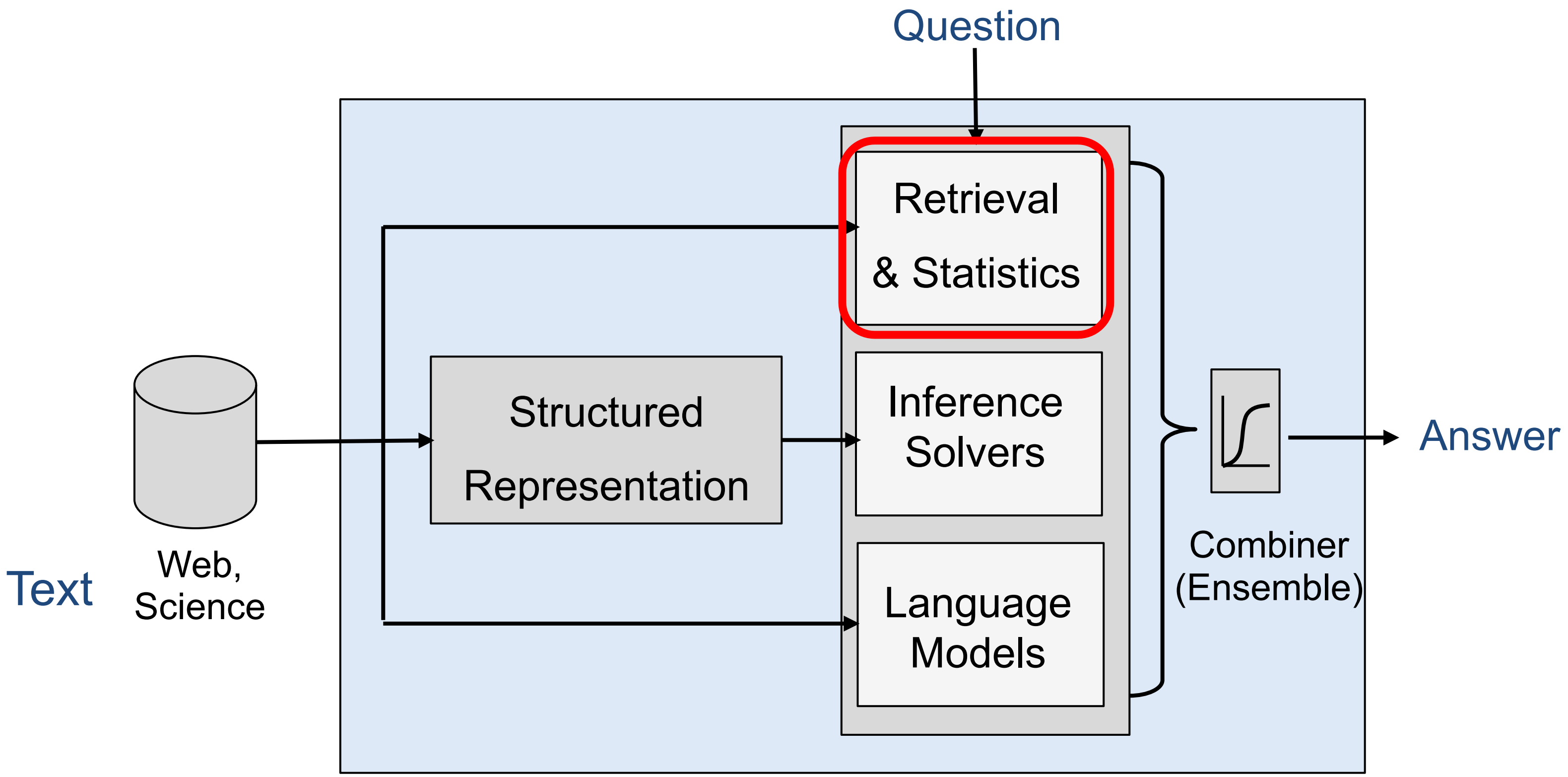
- An ensemble architecture

ARISTO



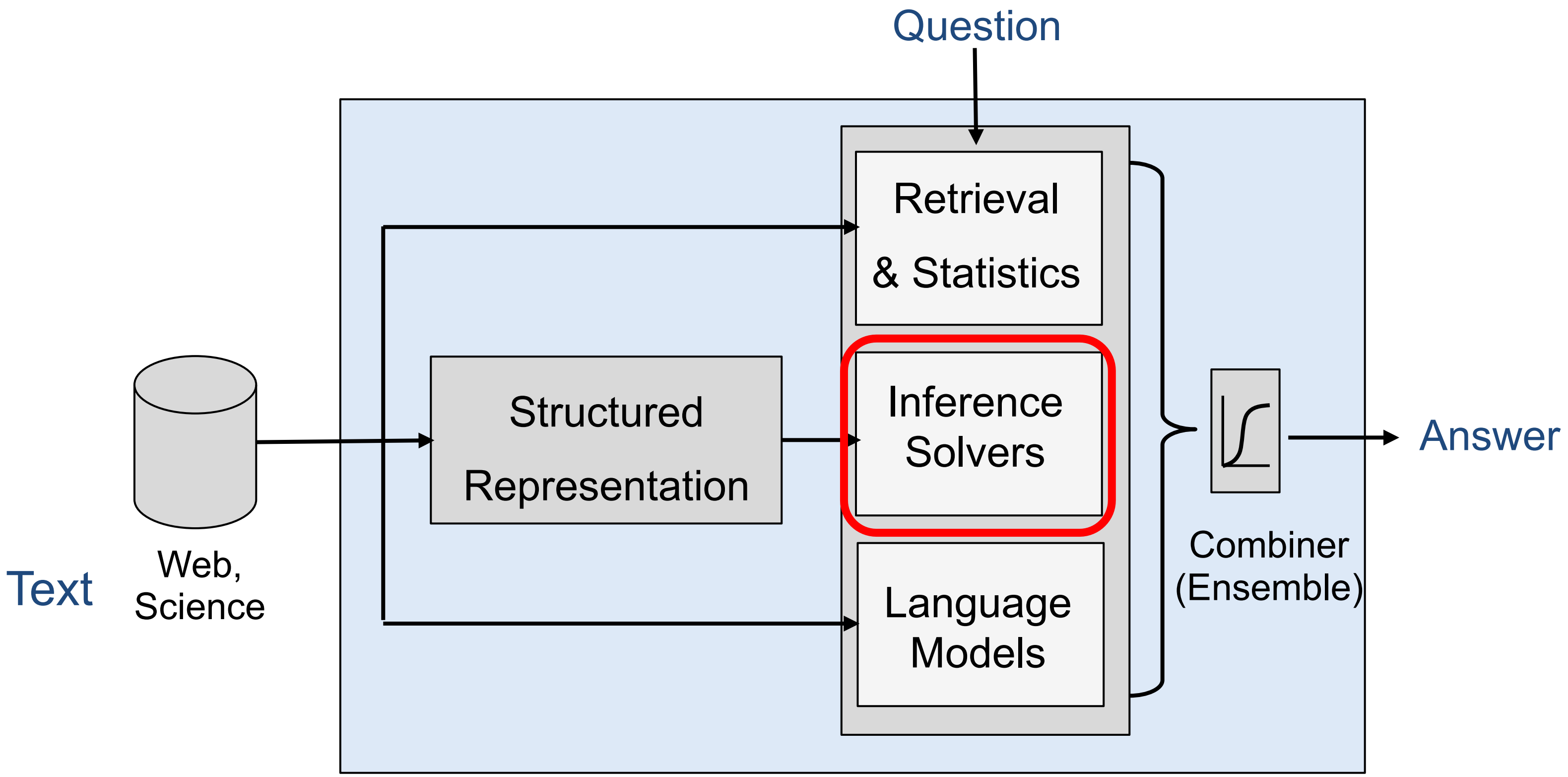
Aristo: an over-simplified overview

- An ensemble architecture



Aristo: an over-simplified overview

- An ensemble architecture



1. Table Knowledge

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

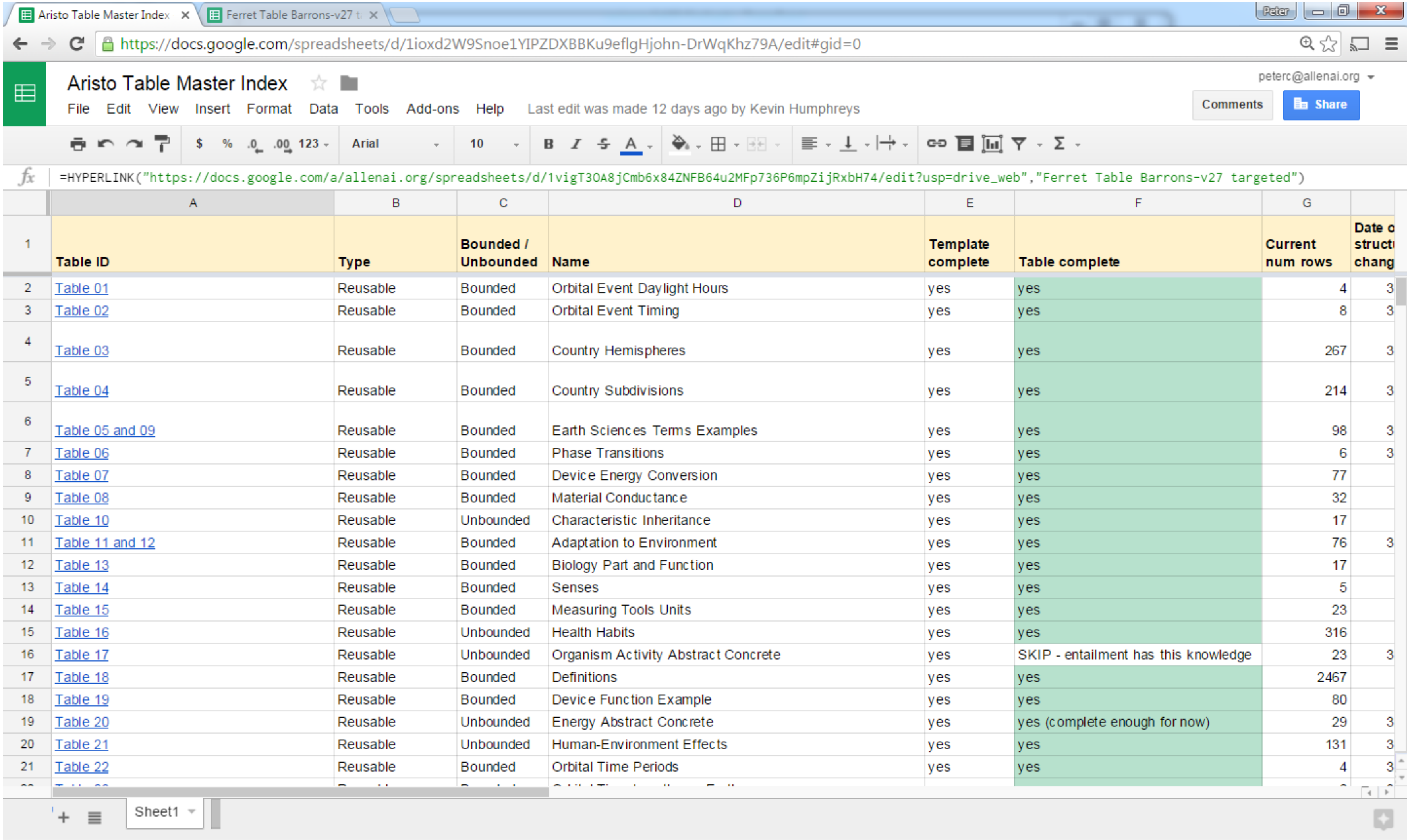
1. Table Knowledge

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

- Daylengths in different months and locations?
- Solstices?
- Where is New York State?
- Which hemisphere is it in?

1. Table Knowledge: Aristo’s Tablestore

- ~120 tables, ~10-500 rows each
- Defined with respect to questions, study guides, syllabus

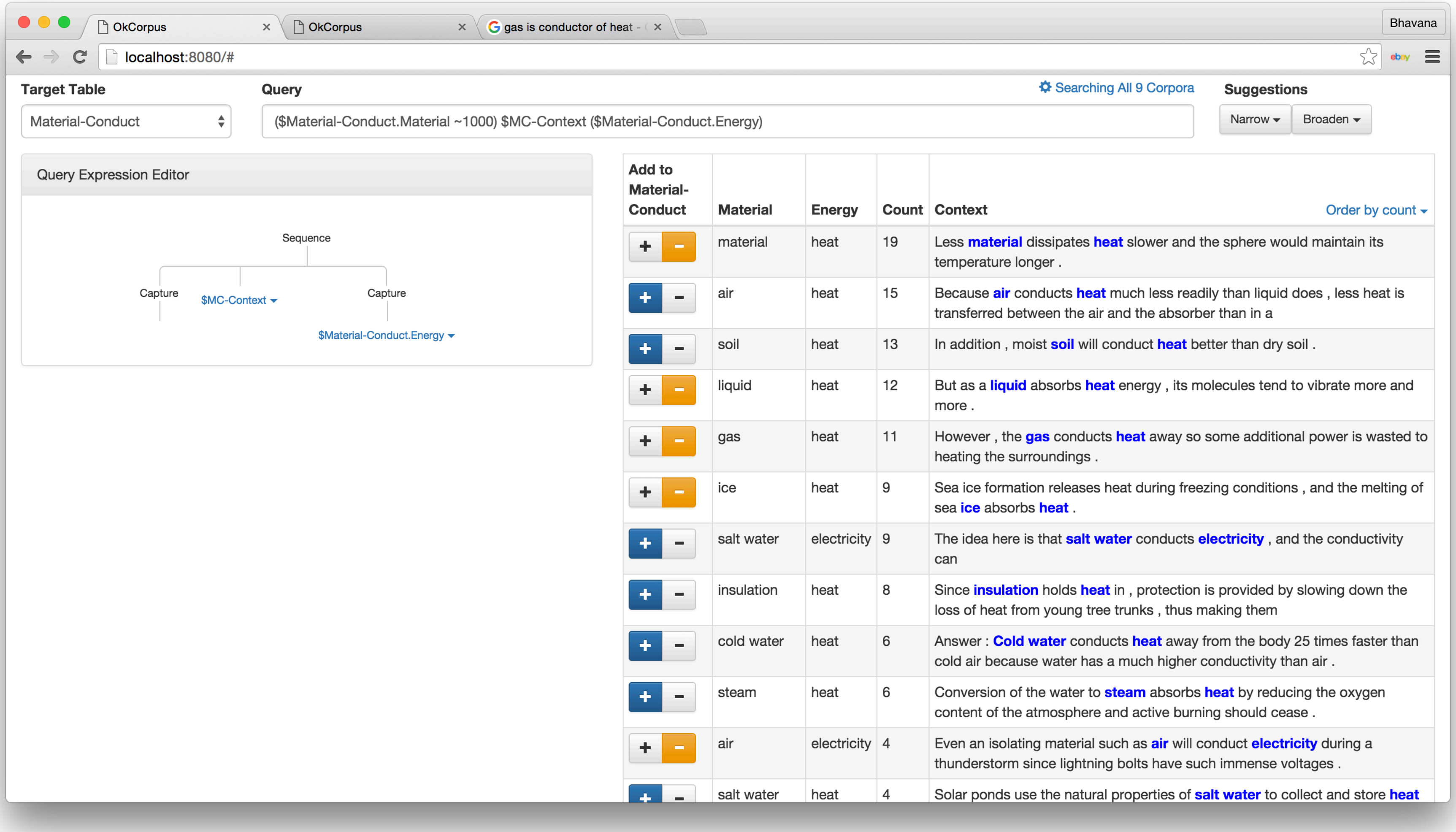


Aristo Table Master Index
Ferret Table Barrons-v27
Peter
https://docs.google.com/spreadsheets/d/1ioxd2W9Snoe1YIPZDXBBKu9eflgHjohn-DrWqKhz79A/edit#gid=0
Aristo Table Master Index
peterc@allenai.org
File
Edit
View
Insert
Format
Data
Tools
Add-ons
Help
Last edit was made 12 days ago by Kevin Humphreys
Comments
Share
Arial
10
Sheet1

=HYPERLINK("https://docs.google.com/a/allenai.org/spreadsheets/d/1vigT30A8jCmb6x84ZNFB64u2MFp736P6mpZijRxbH74/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

IKE – Interactive Knowledge Extraction



Target Table

Material-Conduct

Query

($Material-Conduct.Material ~1000) $MC-Context ($Material-Conduct.Energy)

Searching All 9 Corpora

Suggestions

Query Expression Editor

Add to Material-Conduct	Material	Energy	Count	Context
+ −	material	heat	19	Less **material** dissipates **heat** slower and the sphere would maintain its temperature longer .
+ −	air	heat	15	Because **air** conducts **heat** much less readily than liquid does , less heat is transferred between the air and the absorber than in a
+ −	soil	heat	13	In addition , moist **soil** will conduct **heat** better than dry soil .
+ −	liquid	heat	12	But as a **liquid** absorbs **heat** energy , its molecules tend to vibrate more and more .
+ −	gas	heat	11	However , the **gas** conducts **heat** away so some additional power is wasted to heating the surroundings .
+ −	ice	heat	9	Sea ice formation releases heat during freezing conditions , and the melting of sea **ice** absorbs **heat** .
+ −	salt water	electricity	9	The idea here is that **salt water** conducts **electricity** , and the conductivity can
+ −	insulation	heat	8	Since **insulation** holds **heat** in , protection is provided by slowing down the loss of heat from young tree trunks , thus making them
+ −	cold water	heat	6	Answer : **Cold water** conducts **heat** away from the body 25 times faster than cold air because water has a much higher conductivity than air .
+ −	steam	heat	6	Conversion of the water to **steam** absorbs **heat** by reducing the oxygen content of the atmosphere and active burning should cease .
+ −	air	electricity	4	Even an isolating material such as **air** will conduct **electricity** during a thunderstorm since lightning bolts have such immense voltages .
+ −	salt water	heat	4	Solar ponds use the natural properties of **salt water** to collect and store **heat**

(AKBC’16)

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

1. Table Inference

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

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

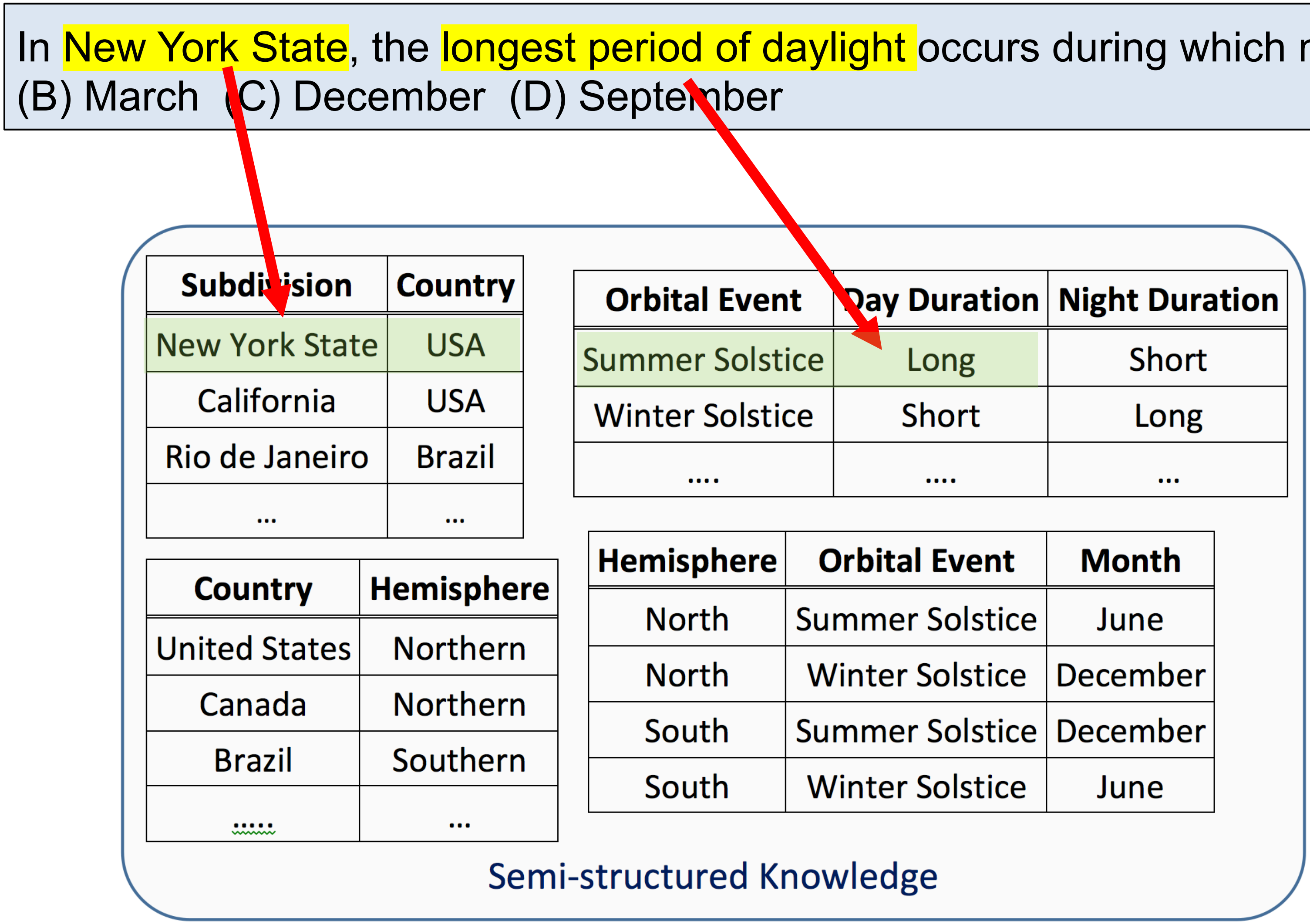
Orbital Event	Day Duration	Night Duration
Summer Solstice	Long	Short
Winter Solstice	Short	Long
….	….	…

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

2. Table Inference



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

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
....	....	...

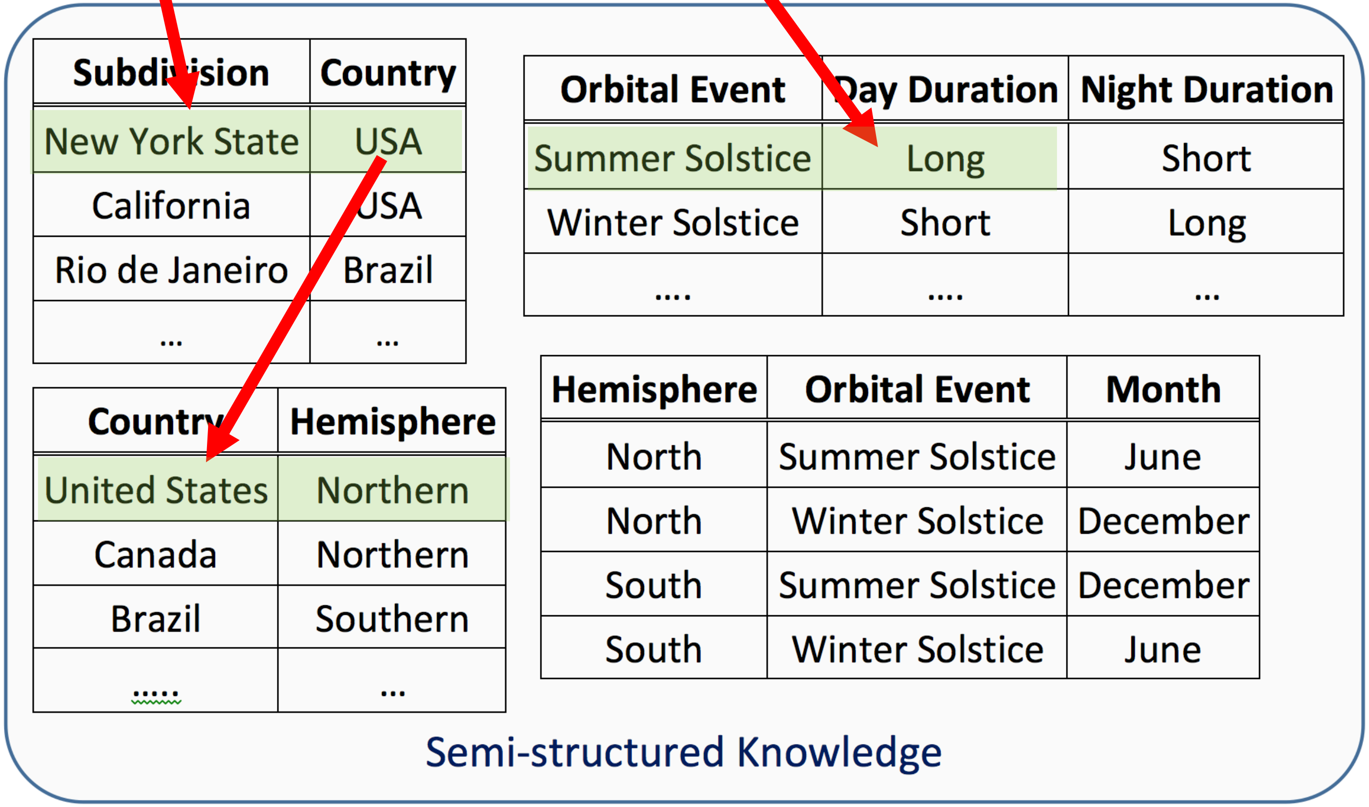
Country	Hemisphere
United States	Northern
Canada	Northern
Brazil	Southern
.....	...

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

Semi-structured Knowledge

1. Table Inference

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



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
….	….	…

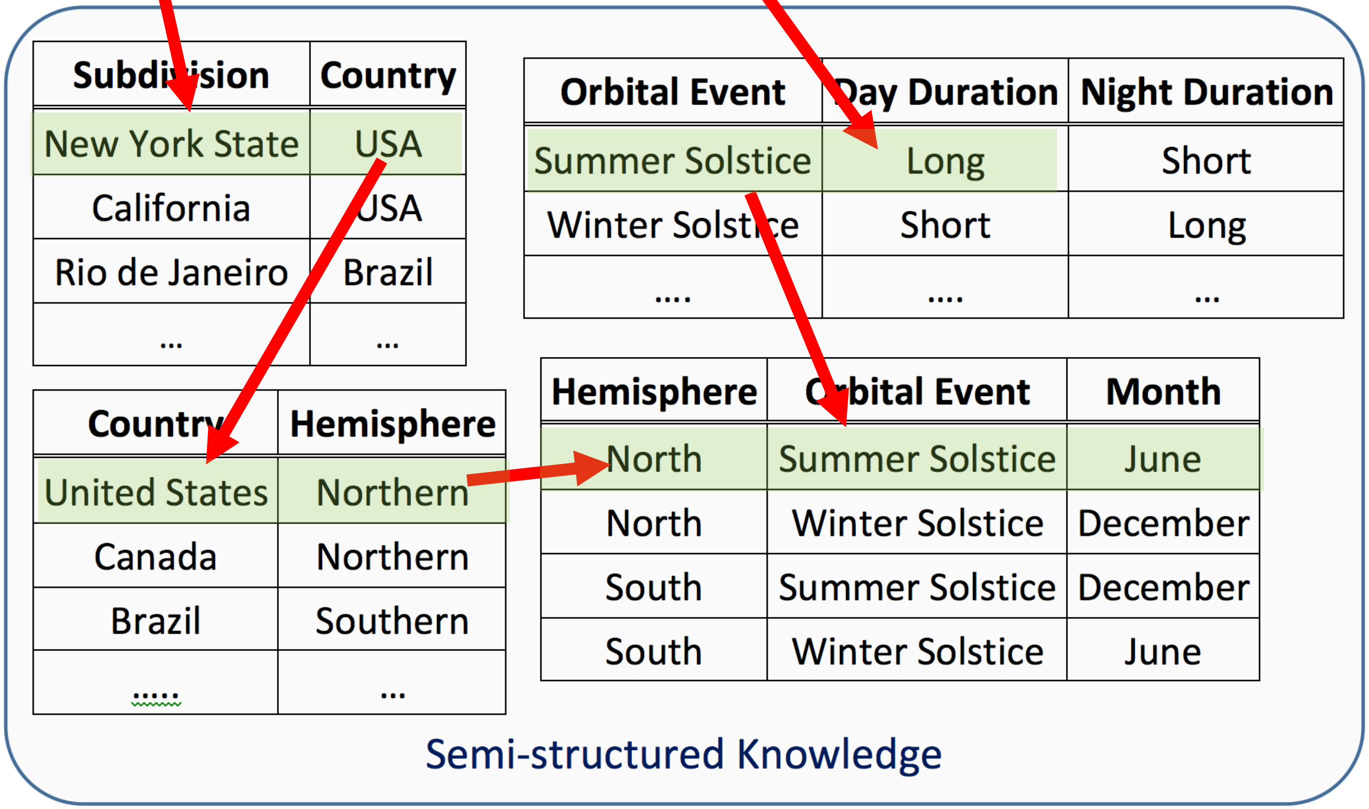
Country	Hemisphere
United States	Northern
Canada	Northern
Brazil	Southern
…..	…

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

Semi-structured Knowledge

1. Table Inference

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



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
….	….	…

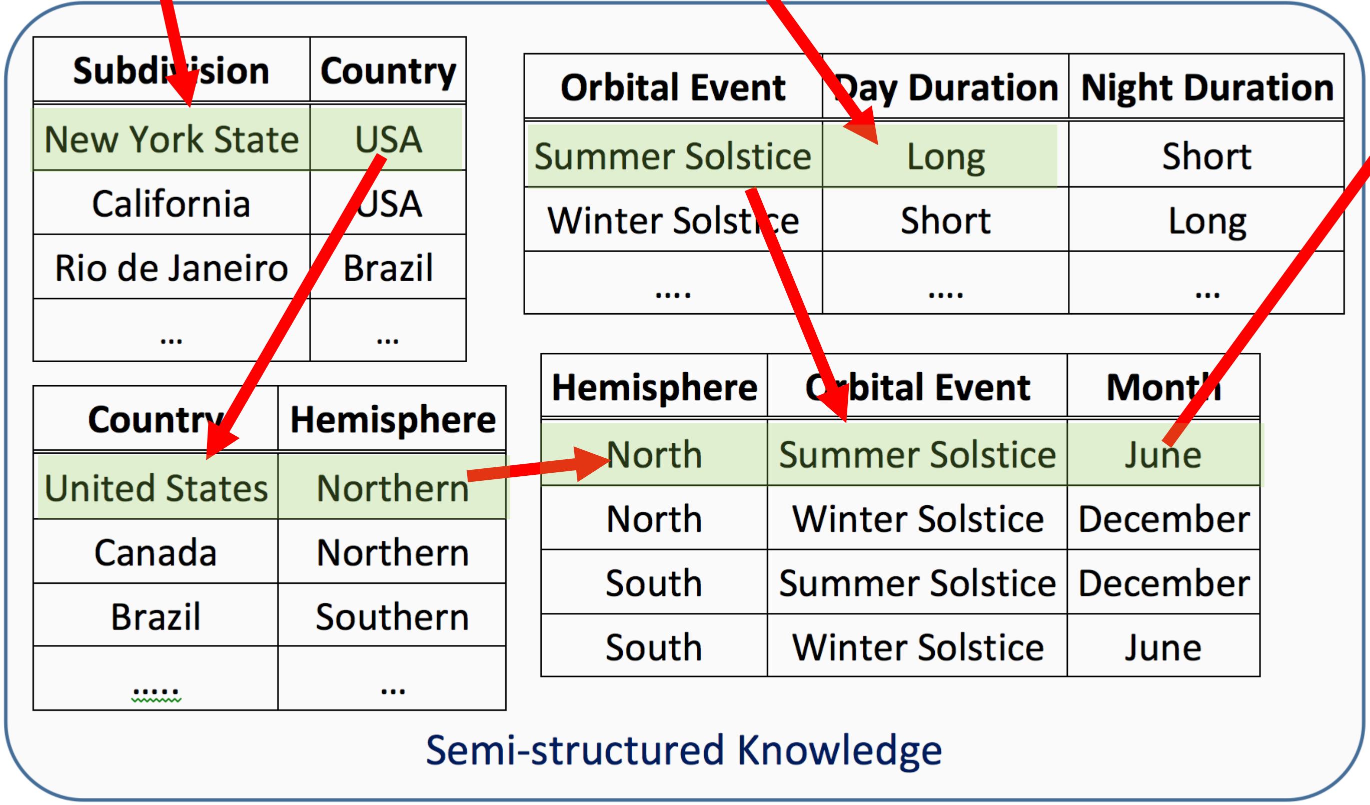
Country	Hemisphere
United States	Northern
Canada	Northern
Brazil	Southern
…..	…

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

Semi-structured Knowledge

1. Table Inference

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



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
….	….	…

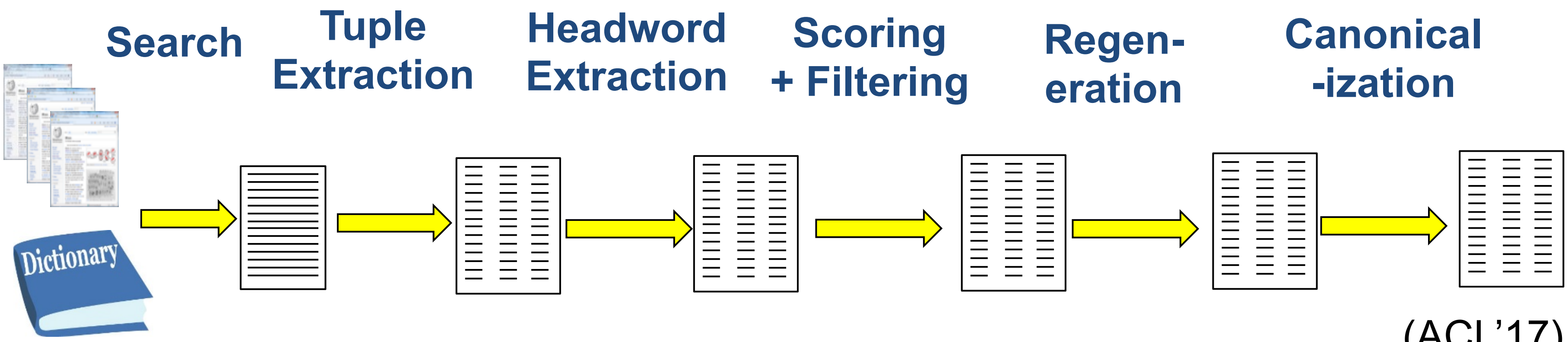
Country	Hemisphere
United States	Northern
Canada	Northern
Brazil	Southern
…..	…

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

Semi-structured Knowledge

IJCAI'16

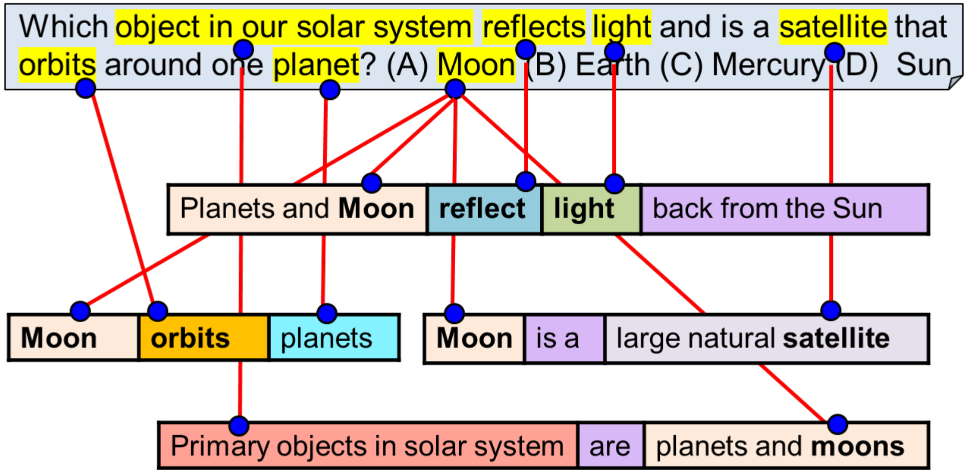
2. Tuple Knowledge



(ACL’17)

Score		Tuple			*Verbalization*
...	...	...	...	...	....
1.00	most	elephant	isa	mammal	*// Elephant isa mammal.*
1.00	most	elephant	isa	pachyderm	*// Elephant isa pachyderm.*
1.00	most	elephant	require	litre water	*// Most elephants require litre water.*
1.00	most	elephant	require	water	*// Most elephants require water.*
...					
0.92	most	elephant	have	curve spine	*// Most elephants have curve spines.*
0.92	most	elephant	need	food	*// Most elephants need food.*
...	...	...	...	...	...
0.83	most	computer	receive	electric energy	*// Most computers receive electric energy.*
0.67	most	computer	solve	problem	*// Most computers solve problems.*
0.60	most	computer	provide	prediction	*// Most computers provide predictions.*
....	...	...	...	...	...

2. Tuple Inference



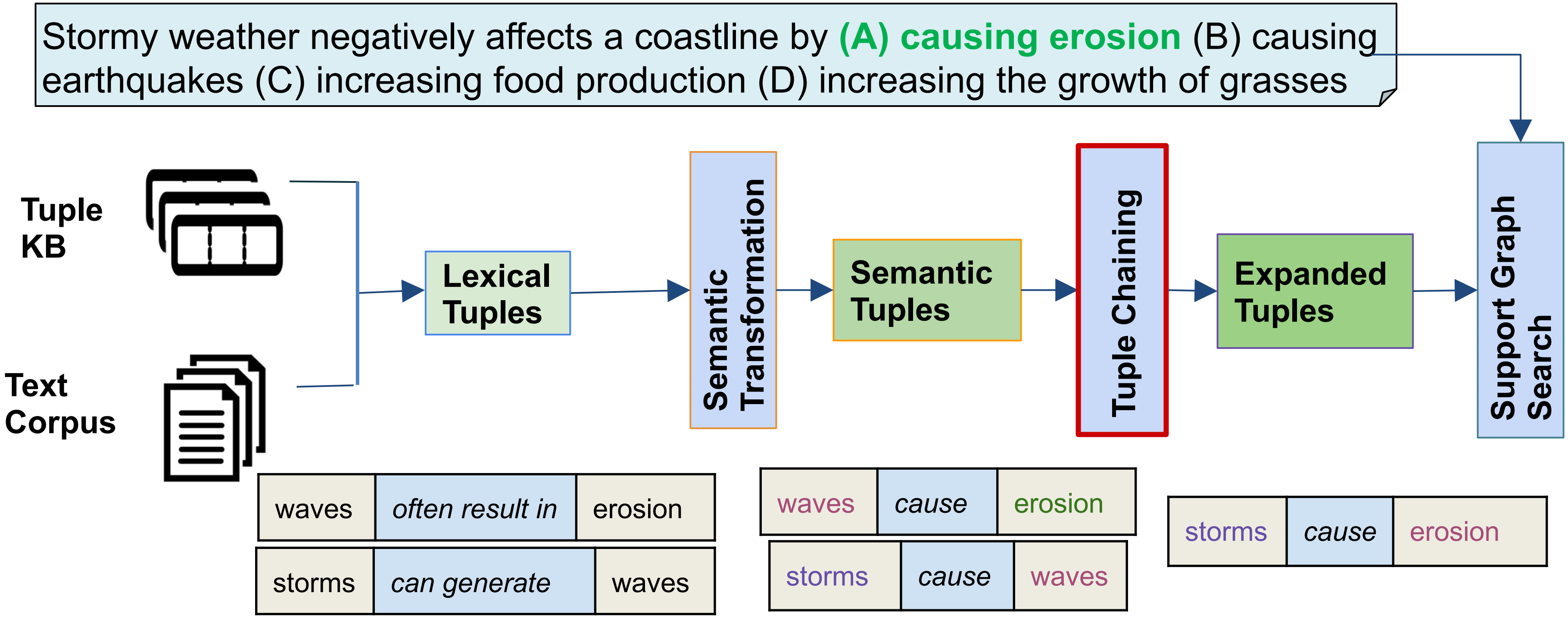
Which object in our solar system reflects light and is a satellite that orbits around one planet? (A) Moon (B) Earth (C) Mercury (D) Sun

Planets and **Moon**	**reflect**	**light**	back from the Sun

Moon	**orbits**	planets

Moon	is a	large natural **satellite**

Primary objects in solar system	are	planets and **moons**



Stormy weather negatively affects a coastline by **(A) causing erosion** (B) causing earthquakes (C) increasing food production (D) increasing the growth of grasses

waves	*often result in*	erosion
storms	*can generate*	waves

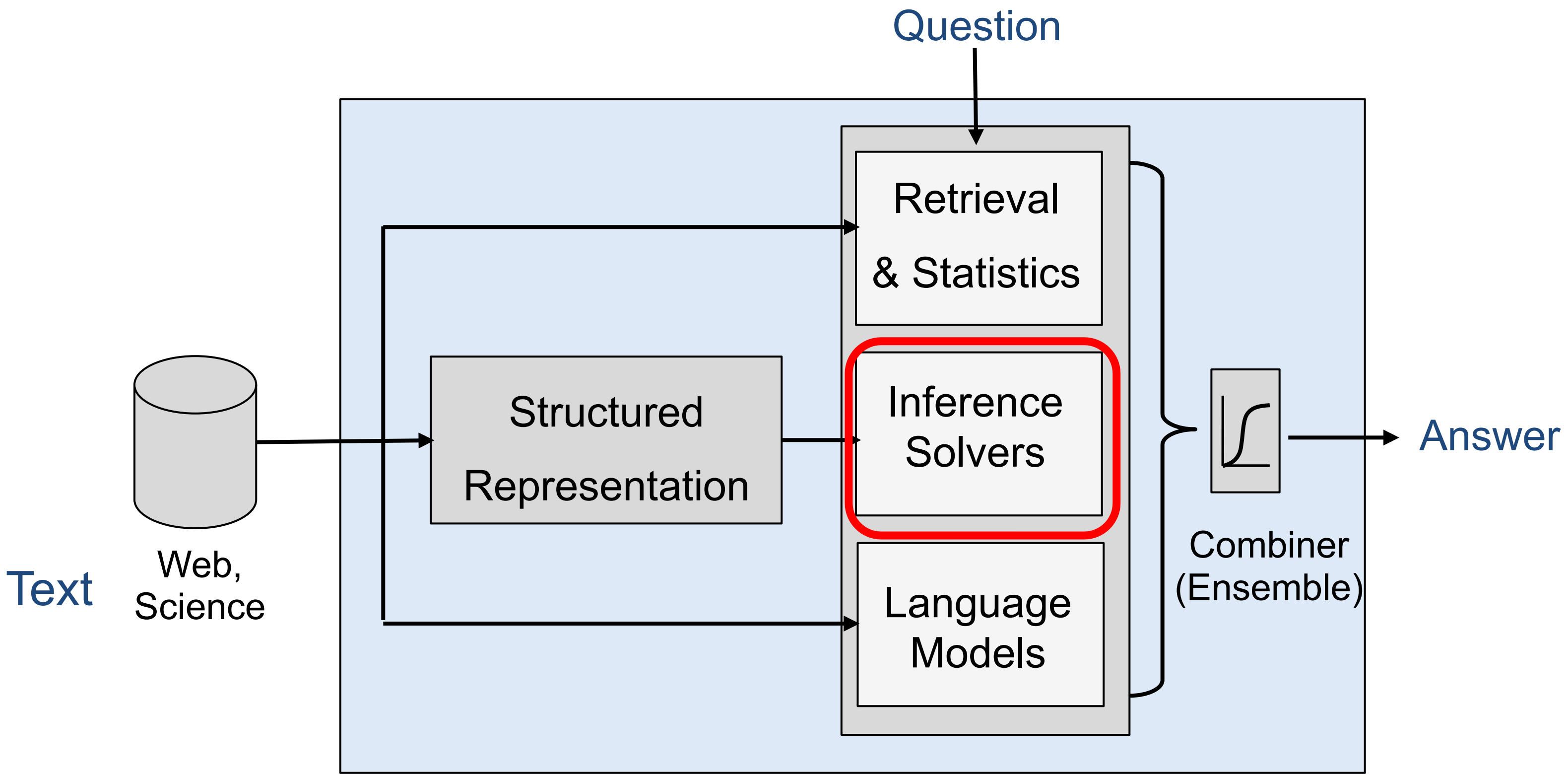
waves	*cause*	erosion
storms	*cause*	waves

storms	*cause*	erosion

Aristo: an over-simplified overview

- An ensemble architecture

ARISTO

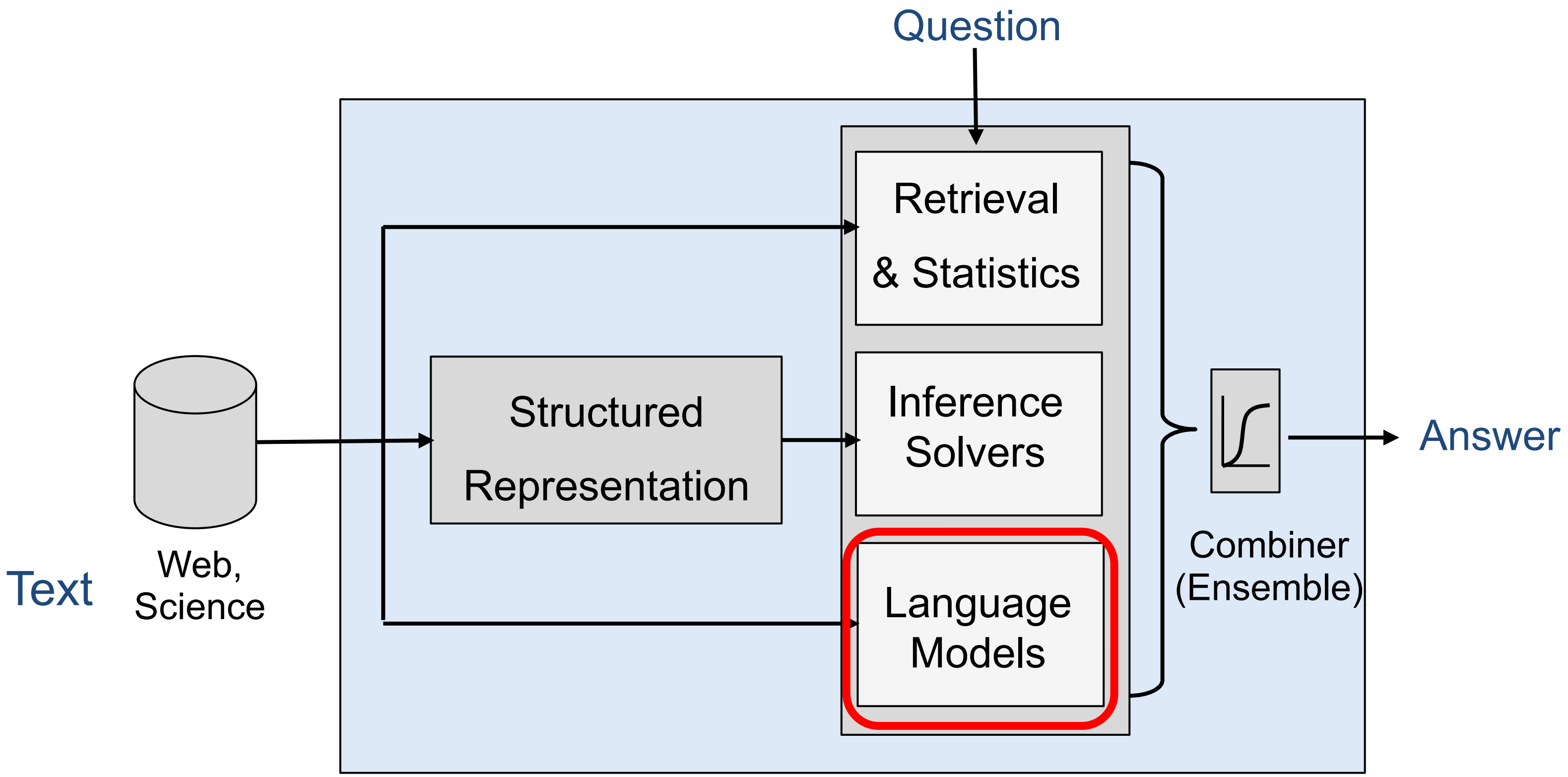


ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

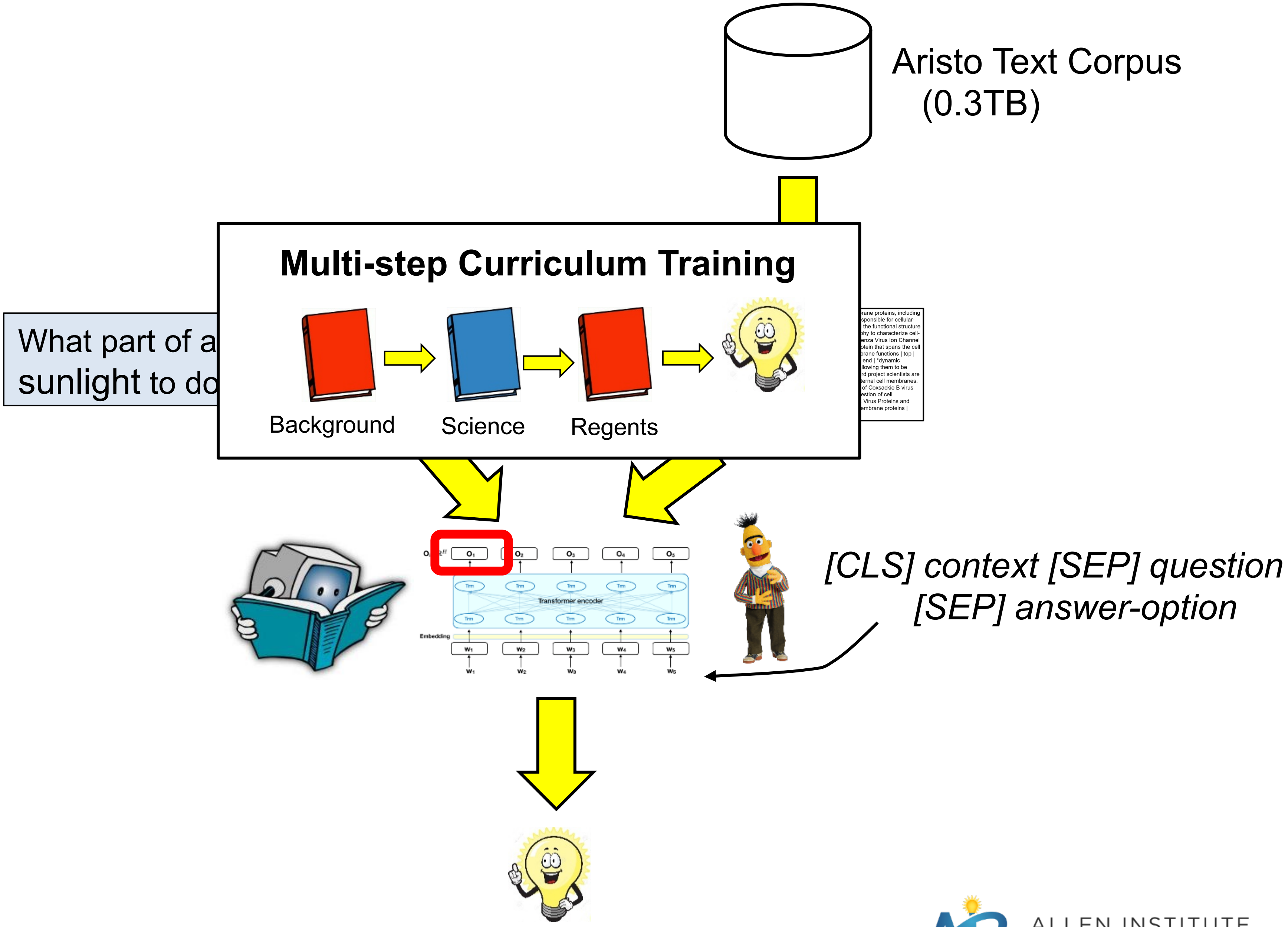
Aristo: an over-simplified overview

- An ensemble architecture

ARISTO



BERT and RoBERTa



Aristo Text Corpus
(0.3TB)

Multi-step Curriculum Training

Background

Science

Regents

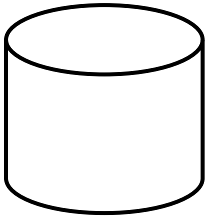
What part of a
sunlight to do

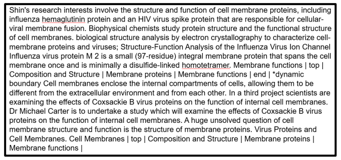
rane proteins, including
sponsible for cellular-
the functional structure
phy to characterize cell-
enza Virus Ion Channel
otein that spans the cell
brane functions | top |
end | *dynamic
llowing them to be
rd project scientists are
ternal cell membranes.
of Coxsackie B virus
estion of cell
Virus Proteins and
embrane proteins |

[CLS] context [SEP] question [SEP] answer-option

ALLEN INSTITUTE
for ARTIFICIAL INTELLIGENCE

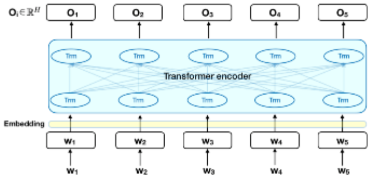
Where is the Knowledge Capture?



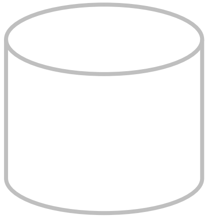


Shin's research interests involve the structure and function of cell membrane proteins, including influenza hemaglutinin protein and an HIV virus spike protein that are responsible for cellular-viral membrane fusion. Biophysical chemists study protein structure and the functional structure of cell membranes. biological structure analysis by electron crystallography to characterize cell-membrane proteins and viruses; Structure-Function Analysis of the Influenza Virus Ion Channel Influenza virus protein M 2 is a small (97-residue) integral membrane protein that spans the cell membrane once and is minimally a disulfide-linked homotetramer. Membrane functions | top | Composition and Structure | Membrane proteins | Membrane functions | end | *dynamic boundary Cell membranes enclose the internal compartments of cells, allowing them to be different from the extracellular environment and from each other. In a third project scientists are examining the effects of Coxsackie B virus proteins on the function of internal cell membranes. Dr Michael Carter is to undertake a study which will examine the effects of Coxsackie B virus proteins on the function of internal cell membranes. A huge unsolved question of cell membrane structure and function is the structure of membrane proteins. Virus Proteins and Cell Membranes. Cell Membranes | top | Composition and Structure | Membrane proteins | Membrane functions |

What part of a plant needs sunlight to do its job? (A) leaf

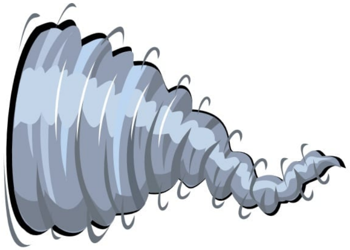


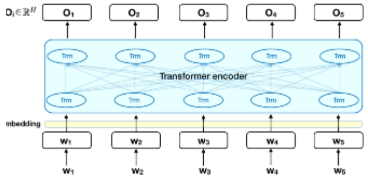
Where is the Knowledge Capture?



What part of a plant needs sunlight to do its job? (A) leaf

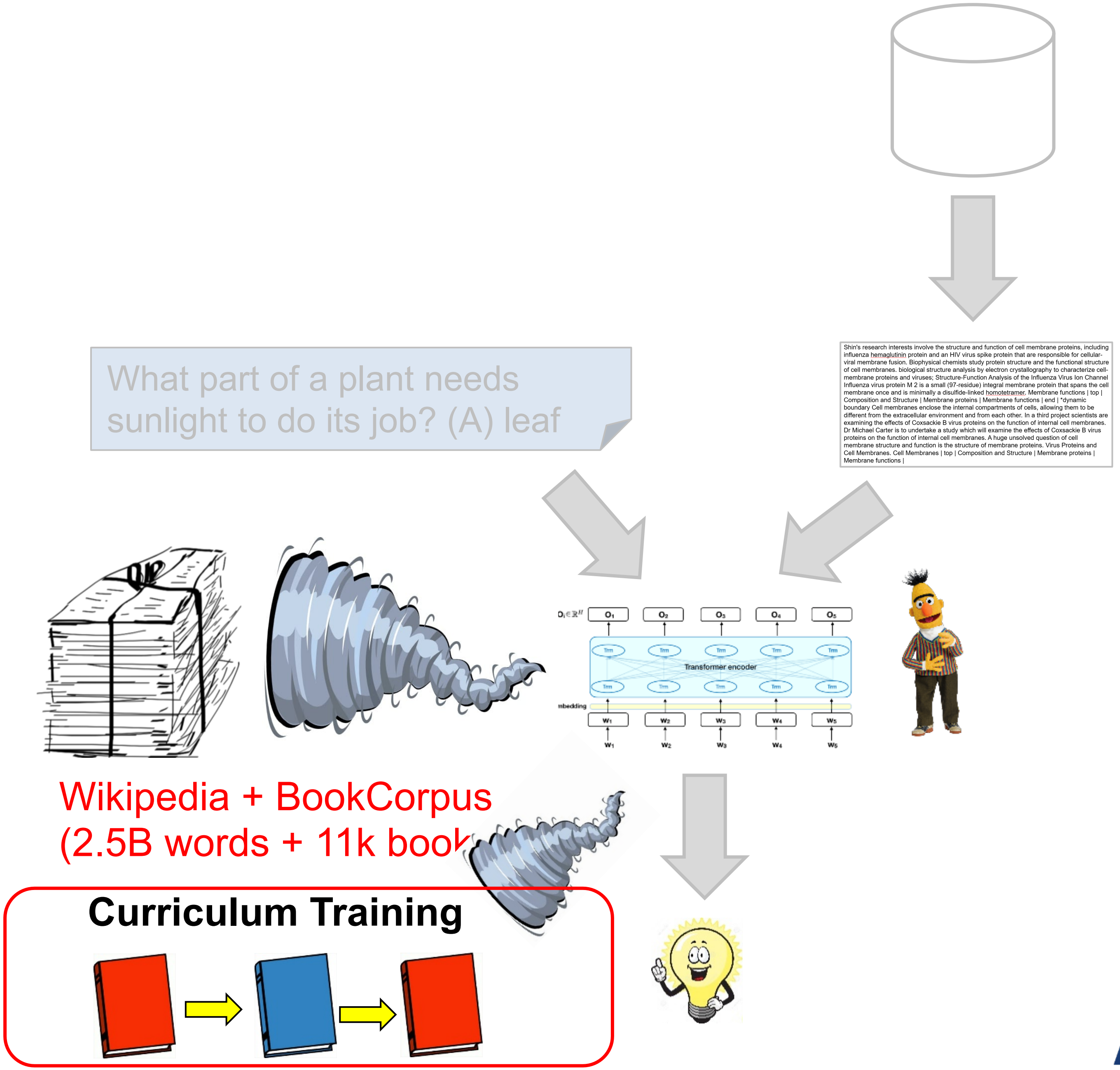
Shin's research interests involve the structure and function of cell membrane proteins, including influenza hemaglutinin protein and an HIV virus spike protein that are responsible for cellular-viral membrane fusion. Biophysical chemists study protein structure and the functional structure of cell membranes. biological structure analysis by electron crystallography to characterize cell-membrane proteins and viruses; Structure-Function Analysis of the Influenza Virus Ion Channel Influenza virus protein M 2 is a small (97-residue) integral membrane protein that spans the cell membrane once and is minimally a disulfide-linked homotetramer. Membrane functions | top | Composition and Structure | Membrane proteins | Membrane functions | end | *dynamic boundary Cell membranes enclose the internal compartments of cells, allowing them to be different from the extracellular environment and from each other. In a third project scientists are examining the effects of Coxsackie B virus proteins on the function of internal cell membranes. Dr Michael Carter is to undertake a study which will examine the effects of Coxsackie B virus proteins on the function of internal cell membranes. A huge unsolved question of cell membrane structure and function is the structure of membrane proteins. Virus Proteins and Cell Membranes. Cell Membranes | top | Composition and Structure | Membrane proteins | Membrane functions |





Wikipedia + BookCorpus
(2.5B words + 11k books)

Where is the Knowledge Capture?



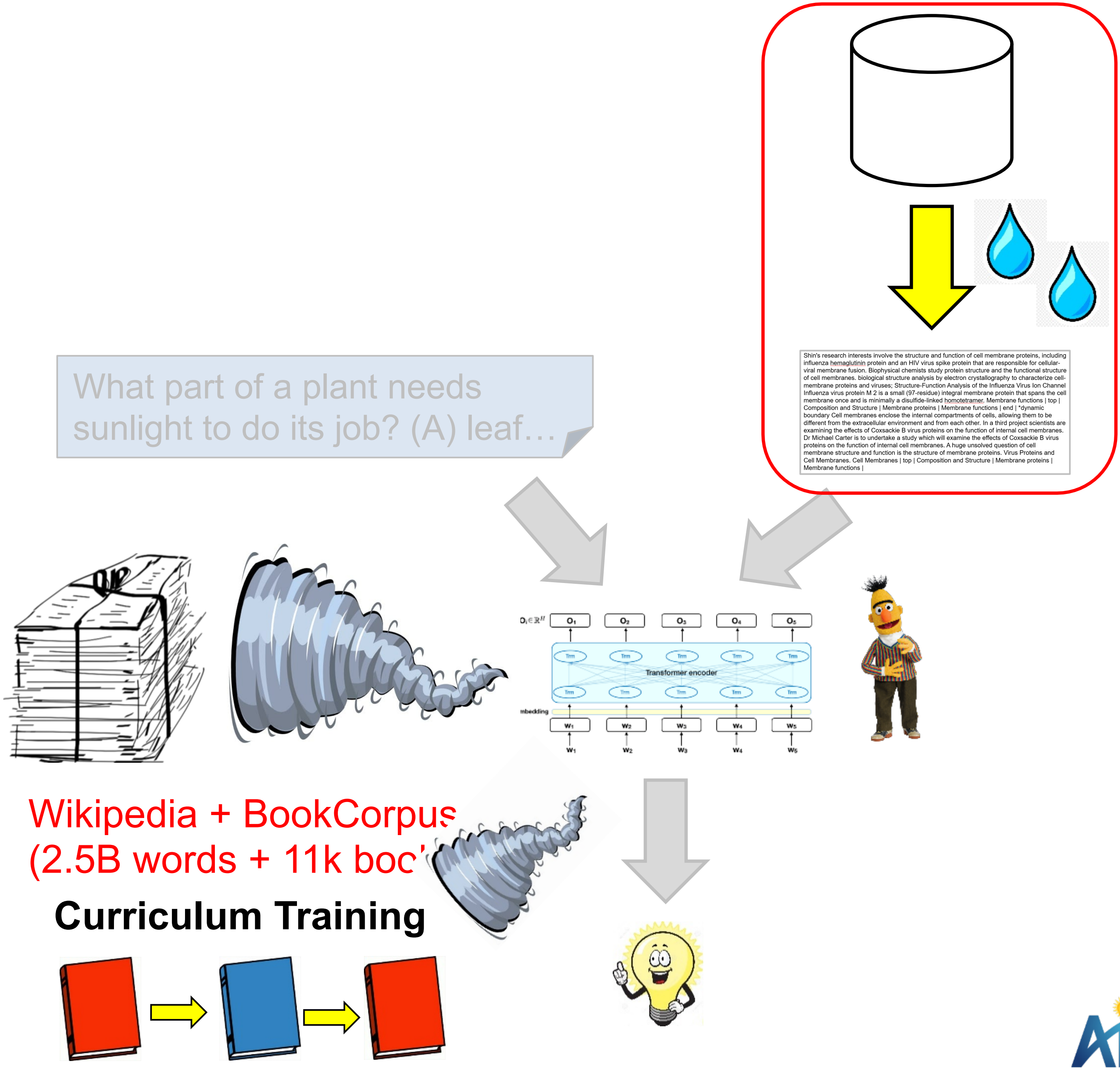
What part of a plant needs sunlight to do its job? (A) leaf

Shin's research interests involve the structure and function of cell membrane proteins, including influenza hemaglutinin protein and an HIV virus spike protein that are responsible for cellular-viral membrane fusion. Biophysical chemists study protein structure and the functional structure of cell membranes. biological structure analysis by electron crystallography to characterize cell-membrane proteins and viruses; Structure-Function Analysis of the Influenza Virus Ion Channel Influenza virus protein M 2 is a small (97-residue) integral membrane protein that spans the cell membrane once and is minimally a disulfide-linked homotetramer. Membrane functions | top | Composition and Structure | Membrane proteins | Membrane functions | end | *dynamic boundary Cell membranes enclose the internal compartments of cells, allowing them to be different from the extracellular environment and from each other. In a third project scientists are examining the effects of Coxsackie B virus proteins on the function of internal cell membranes. Dr Michael Carter is to undertake a study which will examine the effects of Coxsackie B virus proteins on the function of internal cell membranes. A huge unsolved question of cell membrane structure and function is the structure of membrane proteins. Virus Proteins and Cell Membranes. Cell Membranes | top | Composition and Structure | Membrane proteins | Membrane functions |

Wikipedia + BookCorpus
(2.5B words + 11k book

Curriculum Training

Where is the Knowledge Capture?



Aristo Corpus
(2B words)

Shin's research interests involve the structure and function of cell membrane proteins, including influenza hemaglutinin protein and an HIV virus spike protein that are responsible for cellular-viral membrane fusion. Biophysical chemists study protein structure and the functional structure of cell membranes. biological structure analysis by electron crystallography to characterize cell-membrane proteins and viruses; Structure-Function Analysis of the Influenza Virus Ion Channel Influenza virus protein M 2 is a small (97-residue) integral membrane protein that spans the cell membrane once and is minimally a disulfide-linked homotetramer. Membrane functions | top | Composition and Structure | Membrane proteins | Membrane functions | end | *dynamic boundary Cell membranes enclose the internal compartments of cells, allowing them to be different from the extracellular environment and from each other. In a third project scientists are examining the effects of Coxsackie B virus proteins on the function of internal cell membranes. Dr Michael Carter is to undertake a study which will examine the effects of Coxsackie B virus proteins on the function of internal cell membranes. A huge unsolved question of cell membrane structure and function is the structure of membrane proteins. Virus Proteins and Cell Membranes. Cell Membranes | top | Composition and Structure | Membrane proteins | Membrane functions |

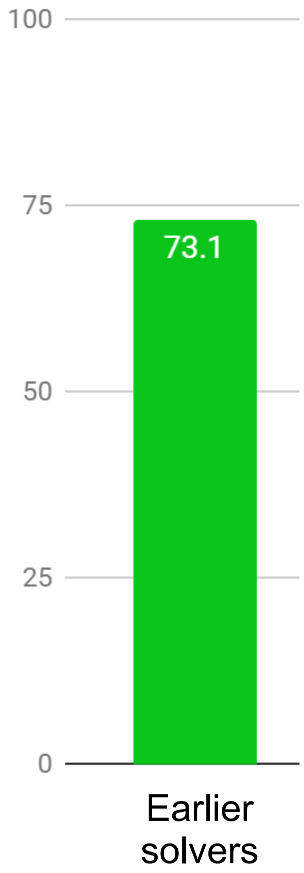
What part of a plant needs sunlight to do its job? (A) leaf…

Wikipedia + BookCorpus
(2.5B words + 11k boc

Curriculum Training

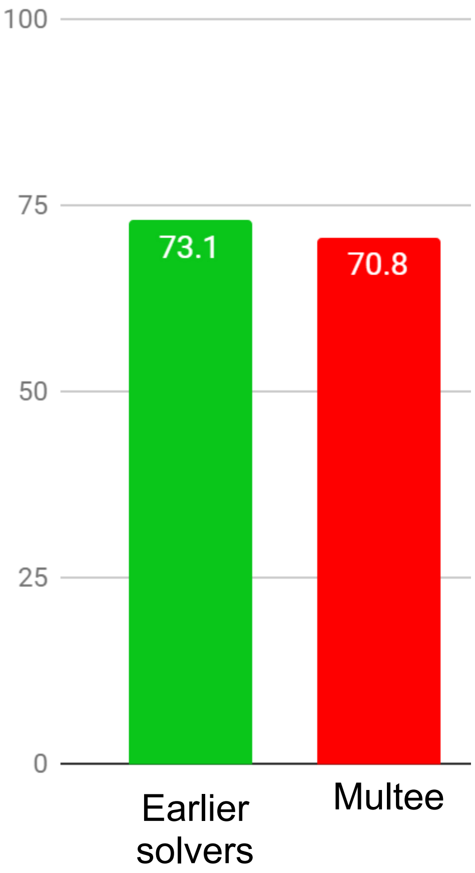
ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

Exploiting Language Models

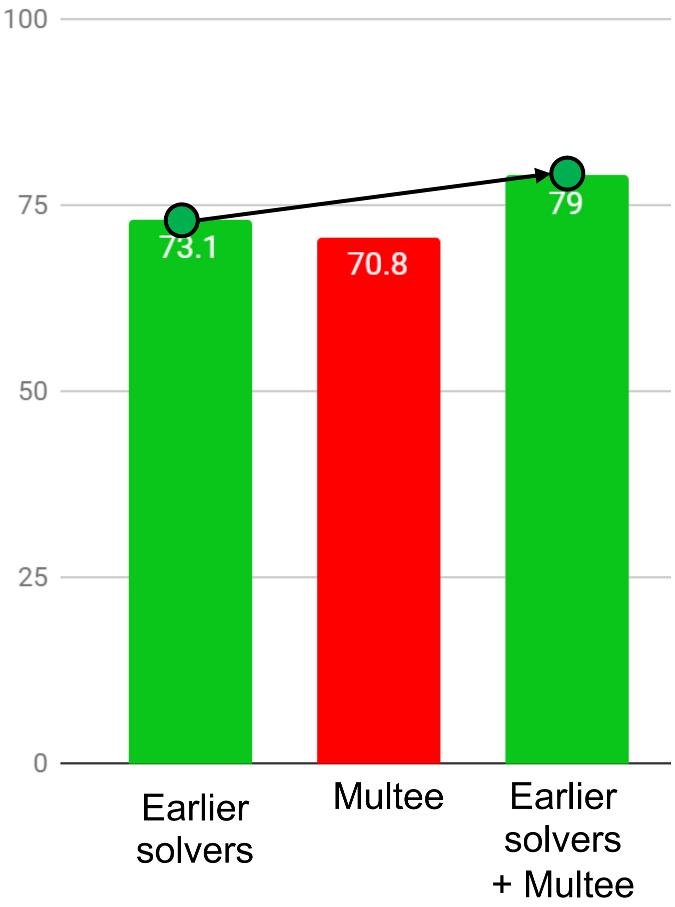


ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

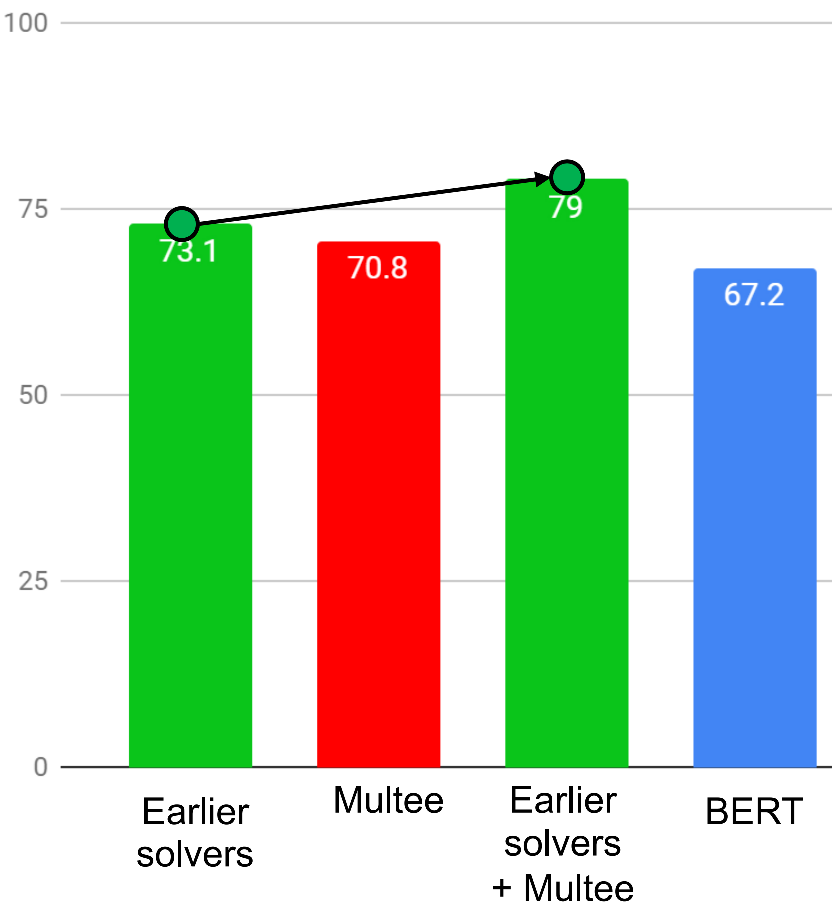
Exploiting Language Models



Exploiting Language Models

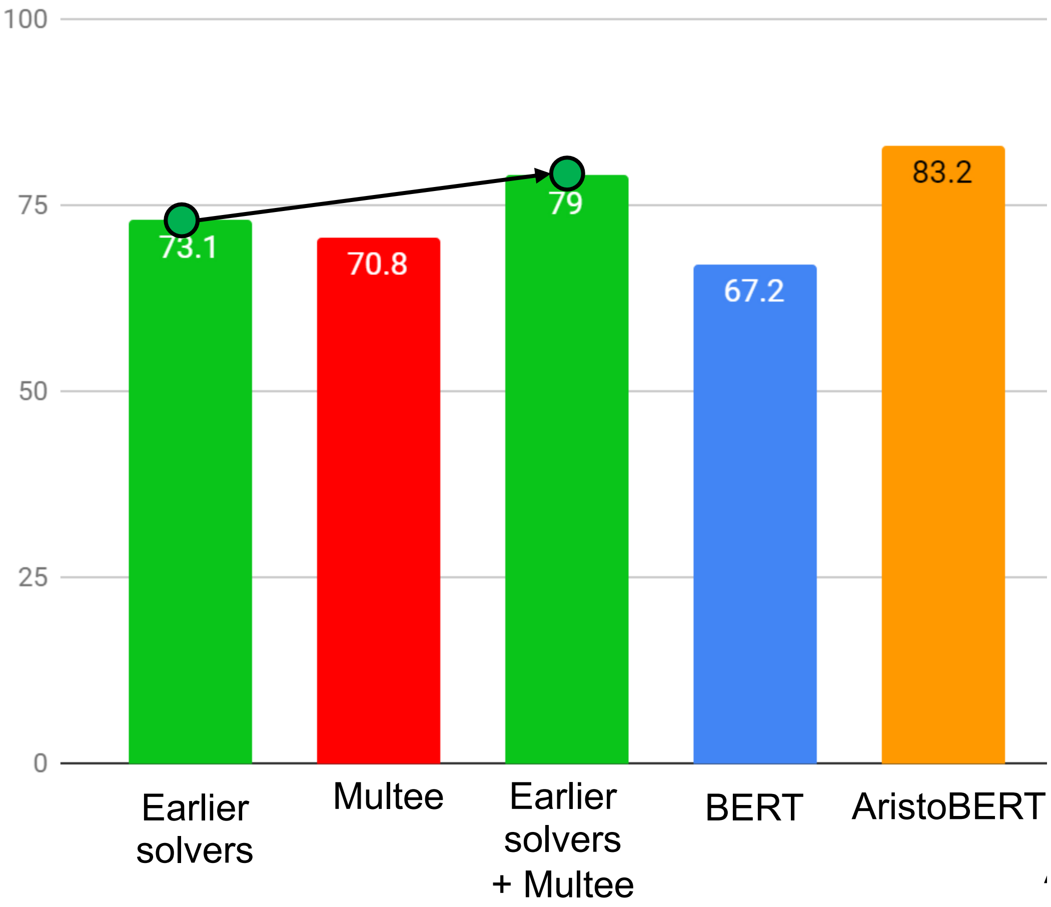


Exploiting Language Models

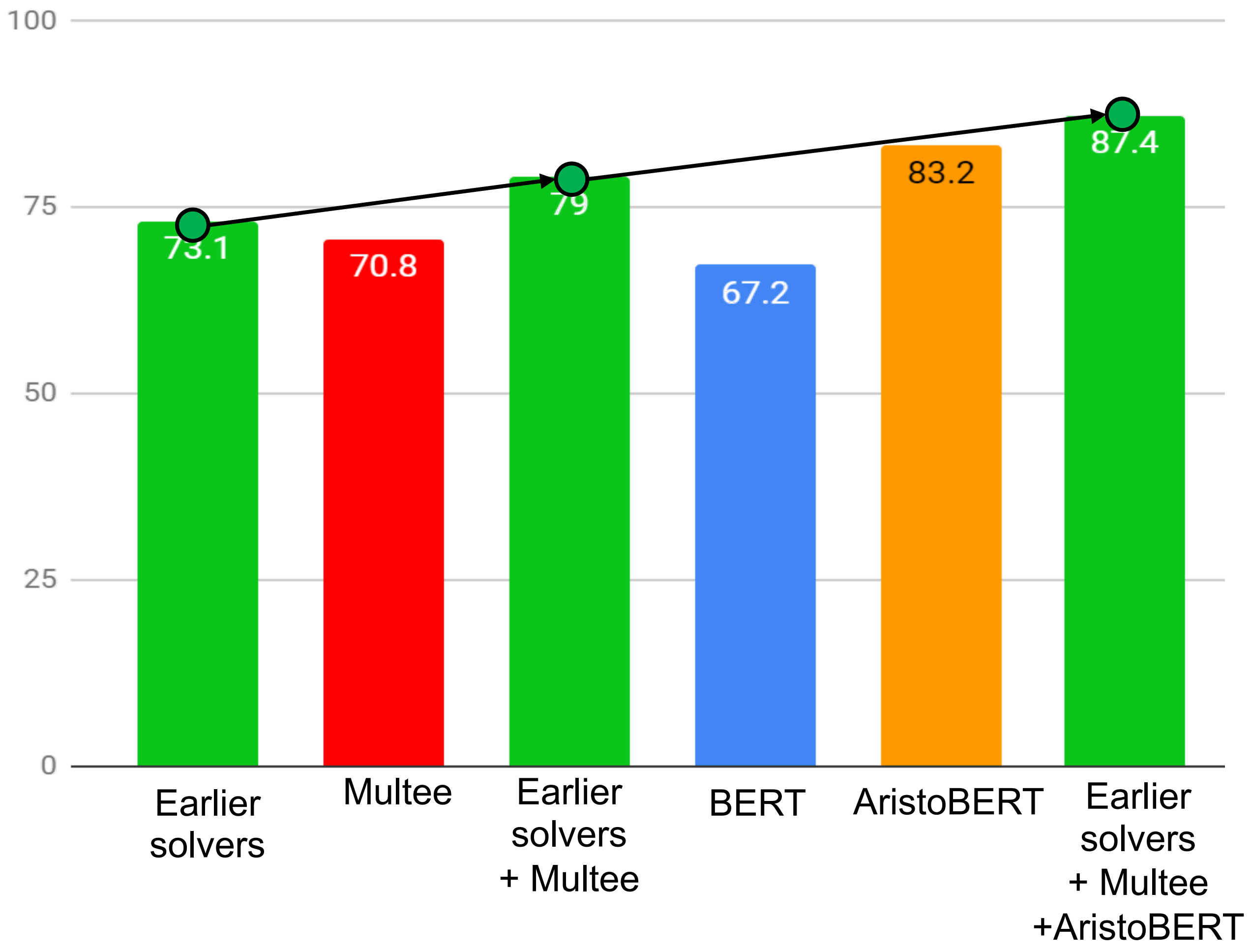


ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

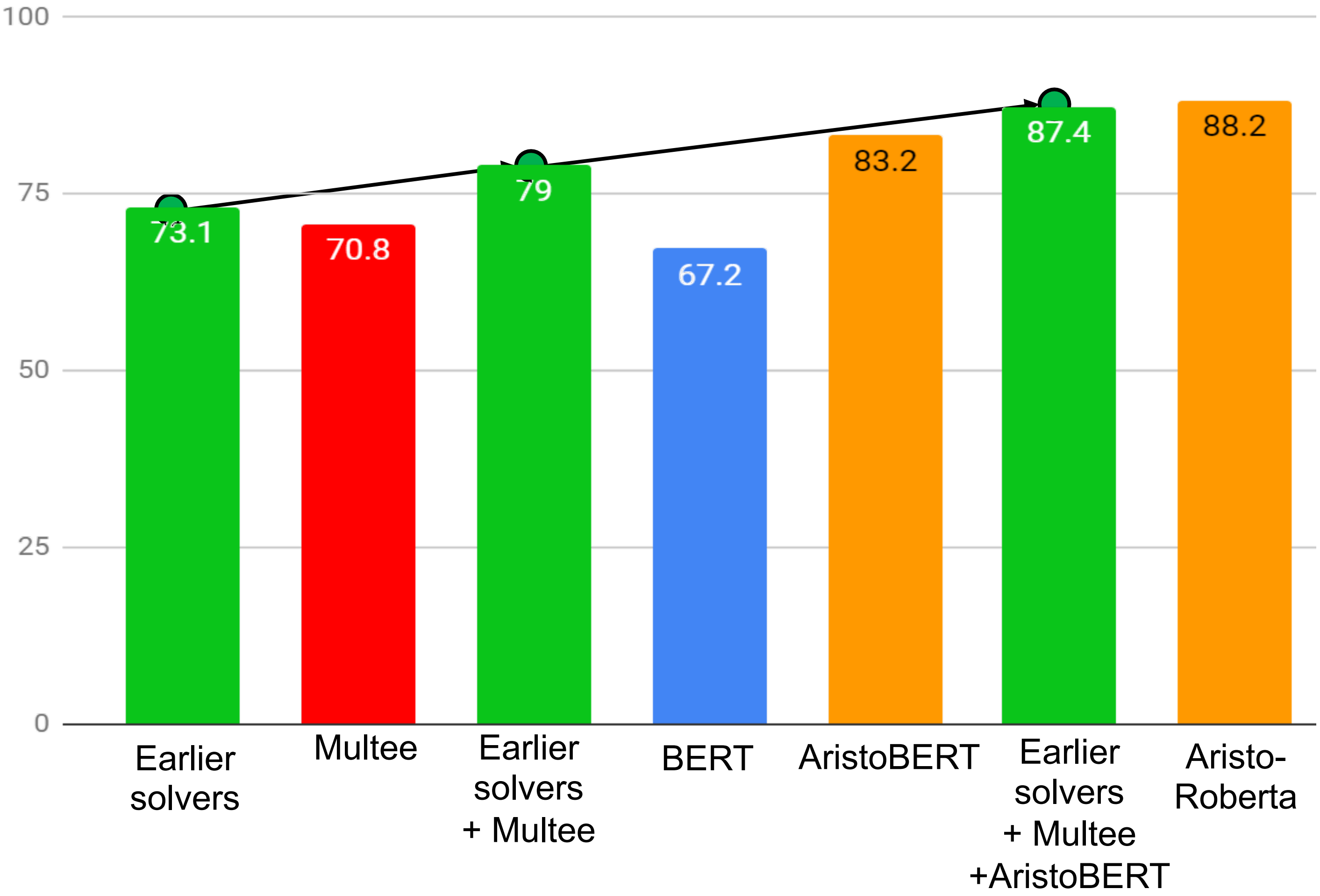
Exploiting Language Models



Exploiting Language Models

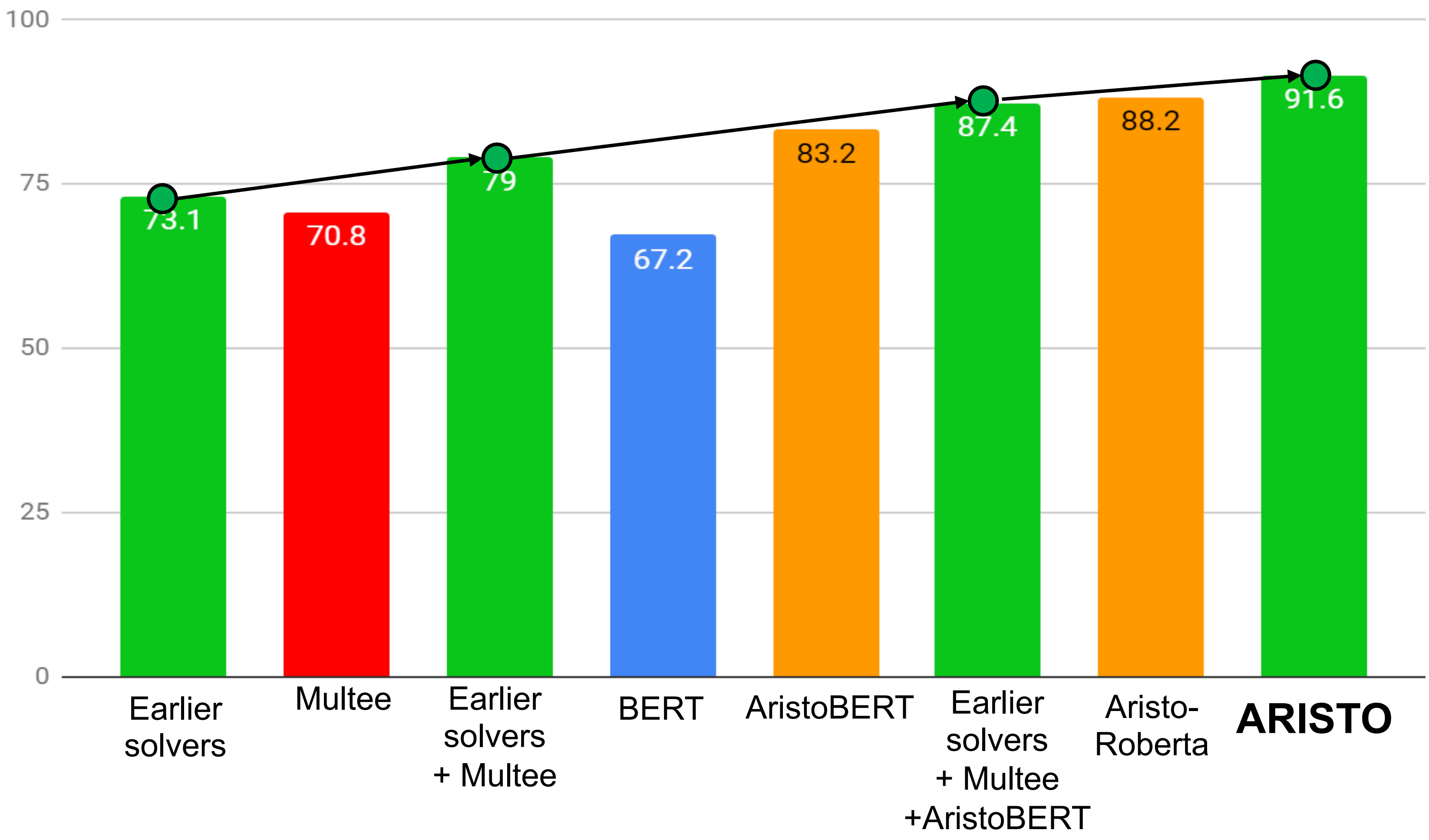


Exploiting Language Models



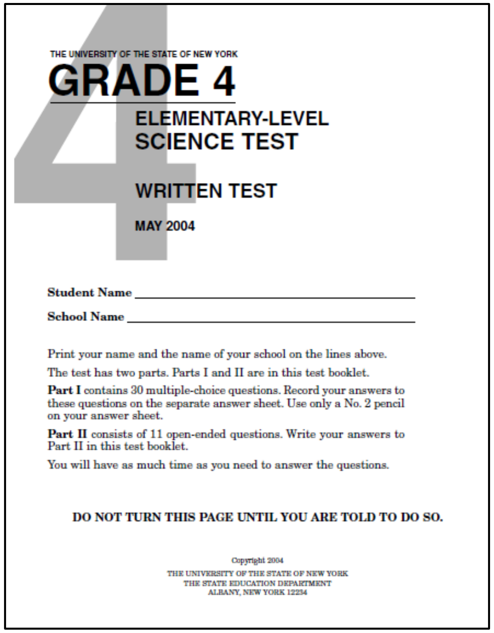
ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

Exploiting Language Models



ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

Similar Progress on 4th Grade NDMC



THE UNIVERSITY OF THE STATE OF NEW YORK

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.

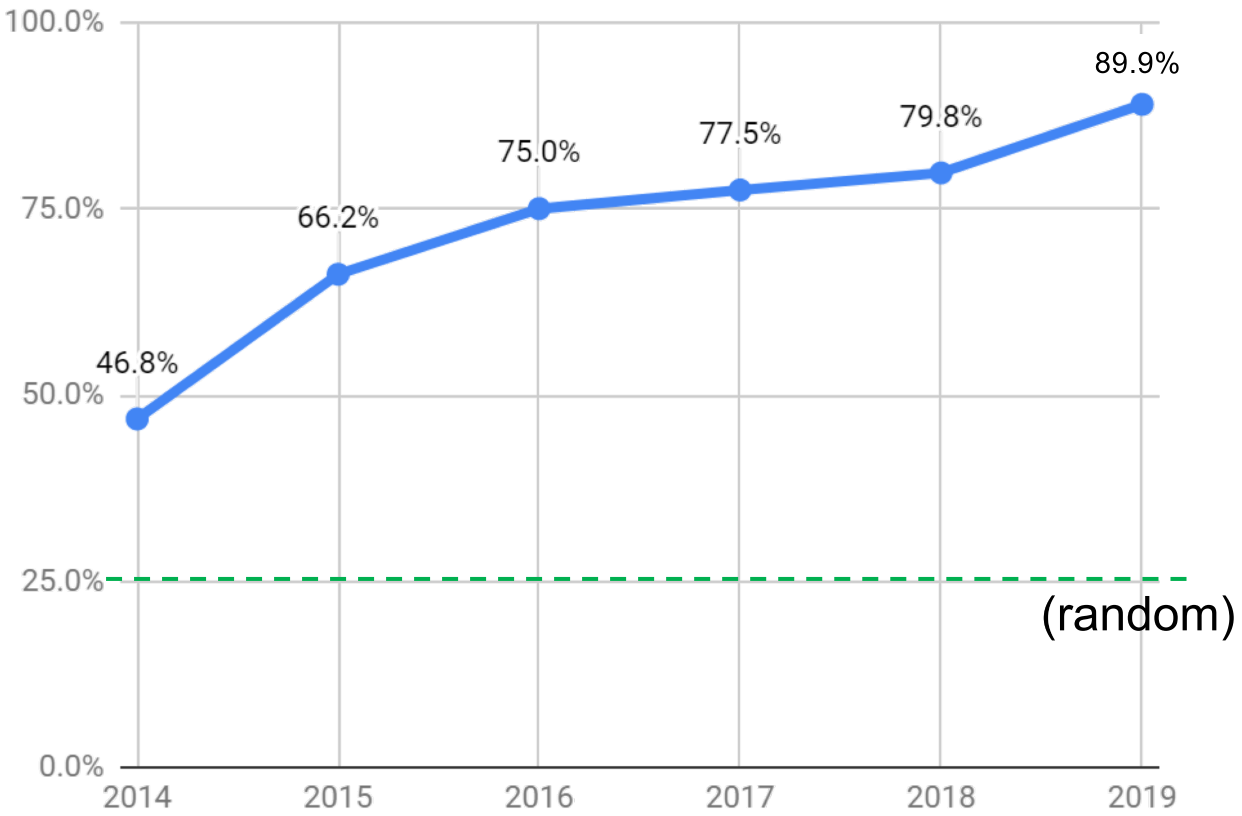
The test has two parts. Parts I and II are in this test booklet.

Part I contains 30 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.

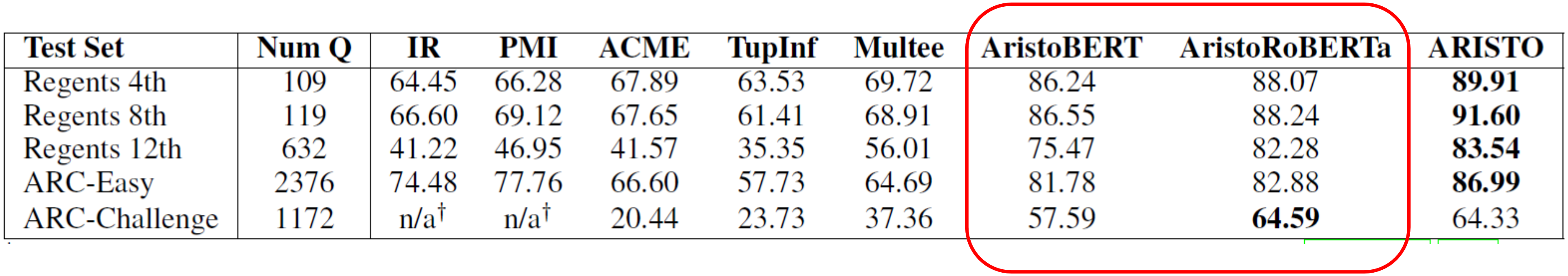


Similarly on 12th grade NDMC:

- Random: 25.0%
- 2014:   40.6%
- 2019:   83.5%

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

Individual Solver Performances



Test Set	Num Q	IR	PMI	ACME	TupInf	Multee	AristoBERT	AristoRoBERTa	ARISTO
Regents 4th	109	64.45	66.28	67.89	63.53	69.72	86.24	88.07	**89.91**
Regents 8th	119	66.60	69.12	67.65	61.41	68.91	86.55	88.24	**91.60**
Regents 12th	632	41.22	46.95	41.57	35.35	56.01	75.47	82.28	**83.54**
ARC-Easy	2376	74.48	77.76	66.60	57.73	64.69	81.78	82.88	**86.99**
ARC-Challenge	1172	n/a[†]	n/a[†]	20.44	23.73	37.36	57.59	**64.59**	64.33

Most of the heavy lifting….

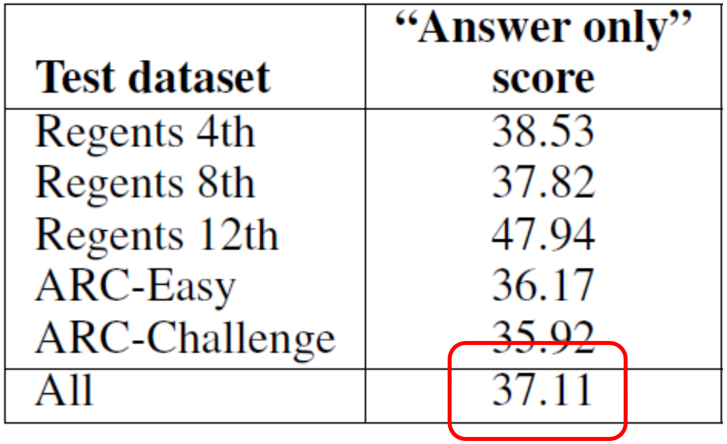
Outline

- Introduction
- How does Aristo work?
- What is going on behind the high scores on the exams?
- Where does Aristo fail?
- What are steps forward?

1. Checking for annotation artifacts

[redacted]

(A) friction
(B) light
(C) force
(D) weather



	“Answer only”
Test dataset	**score**
Regents 4th	38.53
Regents 8th	37.82
Regents 12th	47.94
ARC-Easy	36.17
ARC-Challenge	35.92
All	37.11

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

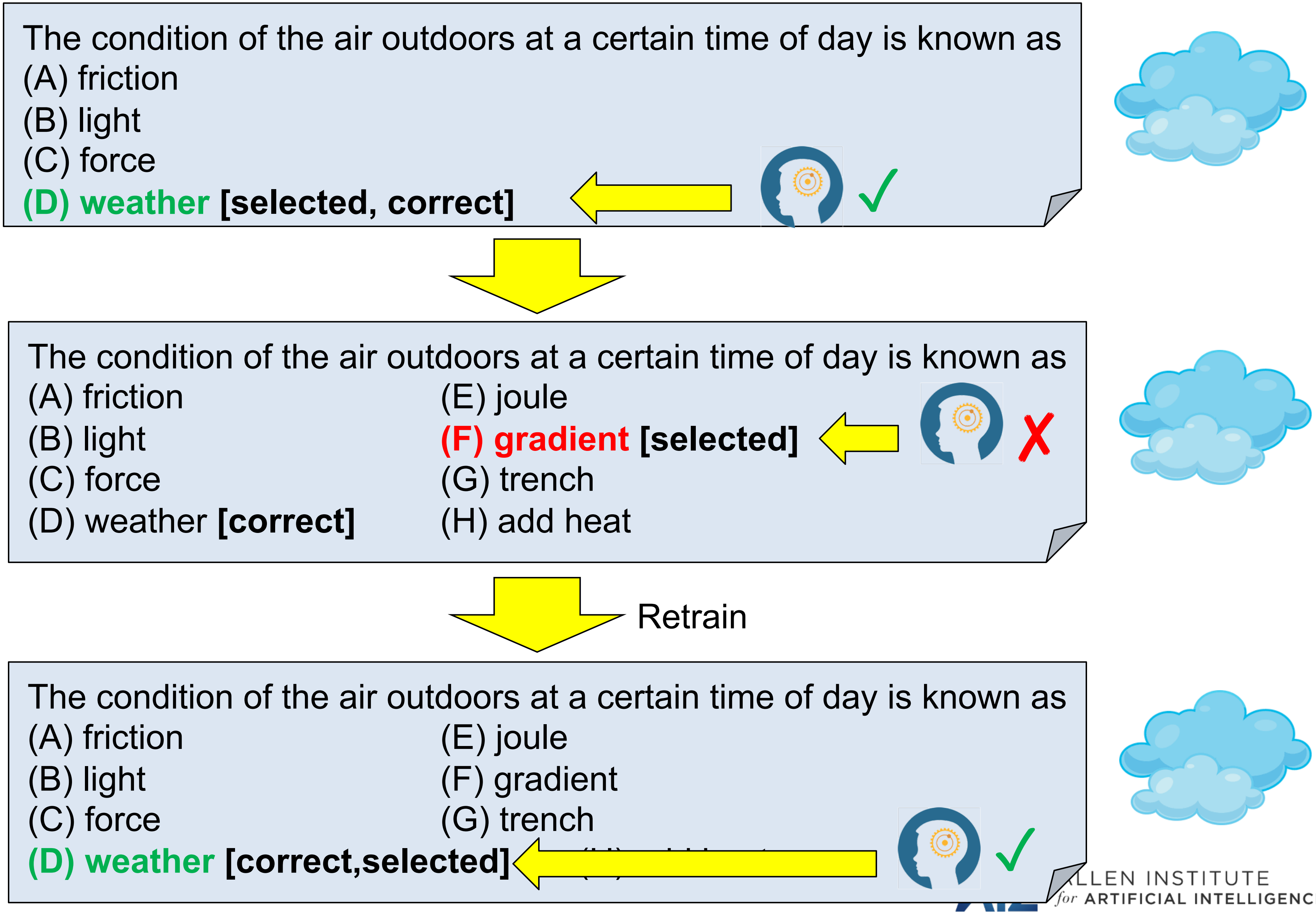
2. Is it fooled by "obviously wrong" answers?

The condition of the air outdoors at a certain time of day is known as
(A) friction
(B) light
(C) force
(D) weather [selected, correct] ✓

The condition of the air outdoors at a certain time of day is known as

(A) friction	(E) joule
(B) light	(F) gradient
(C) force	(G) trench
(D) weather	(H) add heat

2. Is it fooled by “obviously wrong” answers?



The condition of the air outdoors at a certain time of day is known as
(A) friction
(B) light
(C) force
(D) weather [selected, correct] ✓

The condition of the air outdoors at a certain time of day is known as
(A) friction
(B) light
(C) force
(D) weather **[correct]**
(E) joule
(F) gradient [selected] ✗
(G) trench
(H) add heat

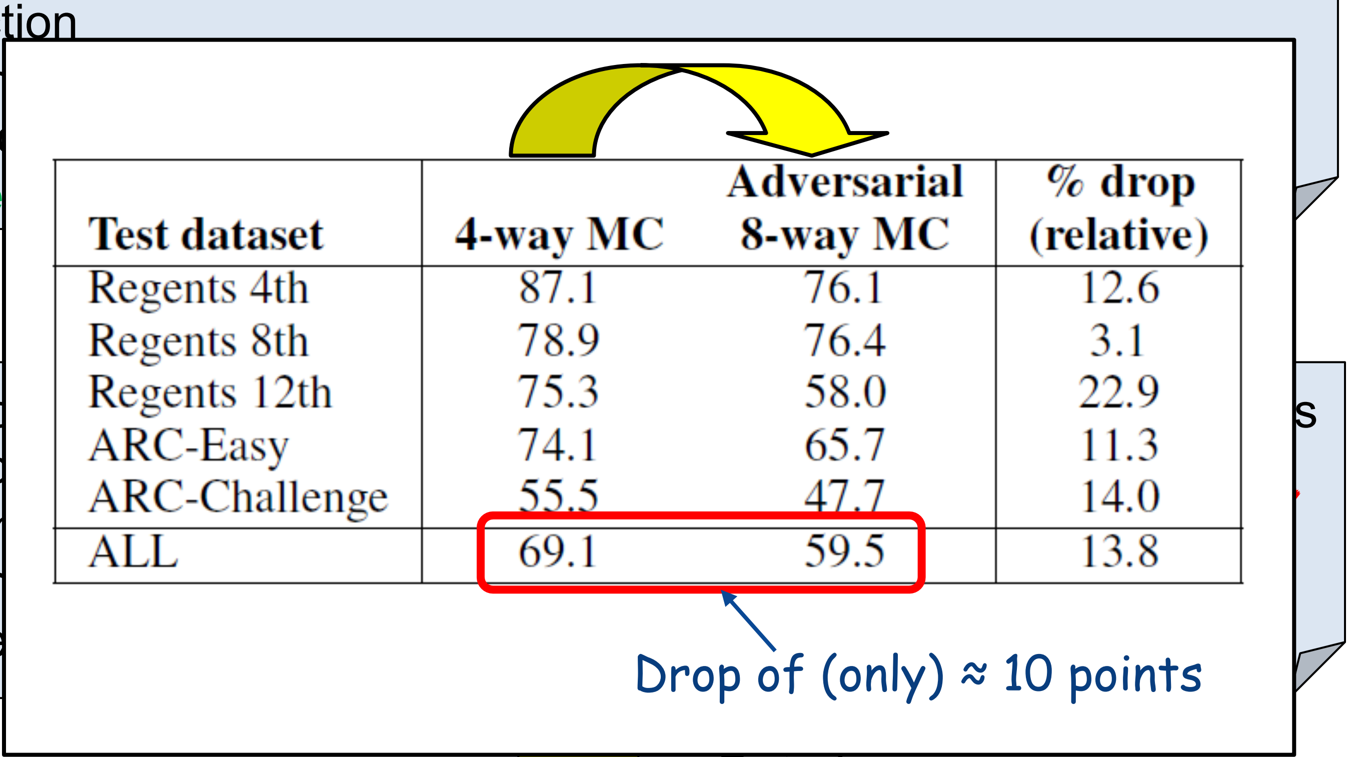
Retrain

The condition of the air outdoors at a certain time of day is known as
(A) friction
(B) light
(C) force
(D) weather [correct,selected] ✓
(E) joule
(F) gradient
(G) trench
(H) add heat

2. Is it fooled by “obviously wrong” answers?

The condition of the air outdoors at a certain time of day is known as
(A) friction
(B) ligh
(C) forc
(D) we

The co ... s
(A) fric
(B) ligh
(C) for
(D) we



Test dataset	4-way MC	Adversarial 8-way MC	% drop (relative)
Regents 4th	87.1	76.1	12.6
Regents 8th	78.9	76.4	3.1
Regents 12th	75.3	58.0	22.9
ARC-Easy	74.1	65.7	11.3
ARC-Challenge	55.5	47.7	14.0
ALL	69.1	59.5	13.8

Drop of (only) ≈ 10 points

Retrain

The condition of the air outdoors at a certain time of day is known as

(A) friction (E) joule
(B) light (F) gradient [selected]
(C) force (G) trench
(D) weather [correct,selected]

✓

ALLEN INSTITUTE
for ARTIFICIAL INTELLIGENCE

3. More than Pattern Matching?

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 ✓

3. More than Pattern Matching?

increasing

raising

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

✓

Which of the following organs does a squirrel ***not*** have
(A) a brain
(B) gills
(C) a heart
(D) lungs

✓

3. More than Pattern Matching?

increasing

raising

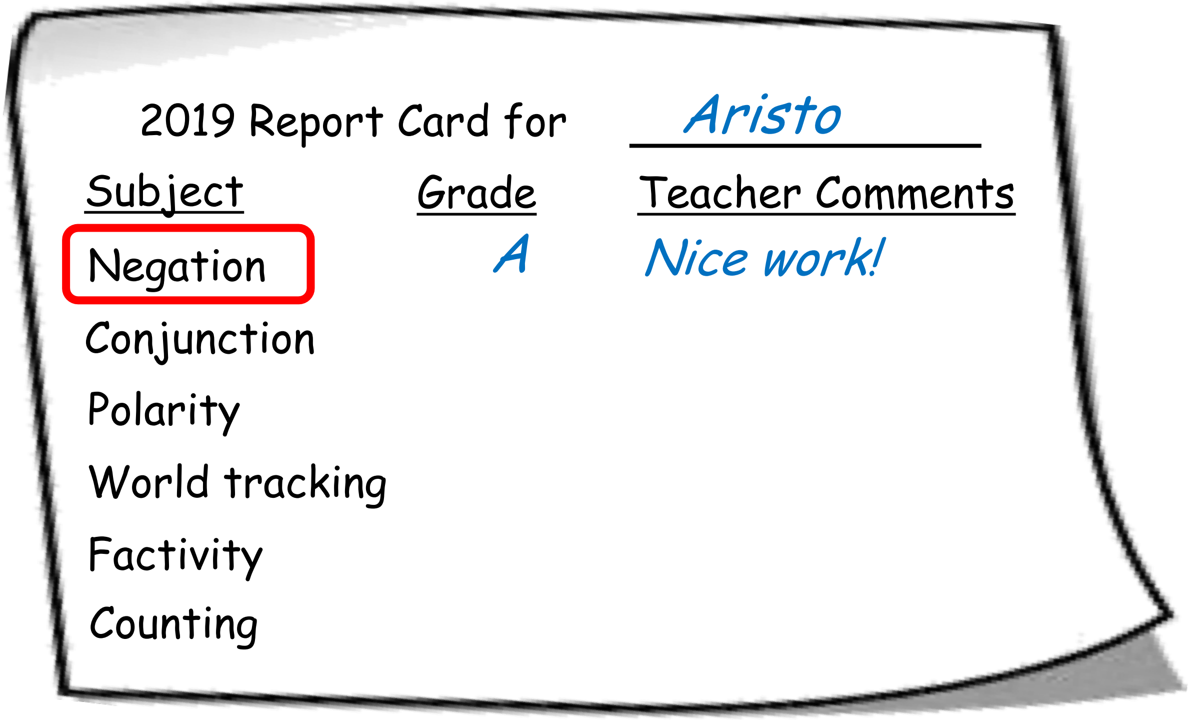
City administrators can encourage energy conservation by
(1) lowering parking fees
(2) building larger parking lots
(3) ~~decreasing~~ the cost of gasoline
(4) ~~lowering~~ the cost of bus and subway fares

✓

Which of the following organs does a squirrel ~~***not***~~ have
(A) a brain
(B) gills
(C) a heart
(D) lungs

✓

3. More than Pattern Matching?



2019 Report Card for Aristo

Subject	Grade	Teacher Comments
Negation	A	Nice work!
Conjunction		
Polarity		
World tracking		
Factivity		
Counting		

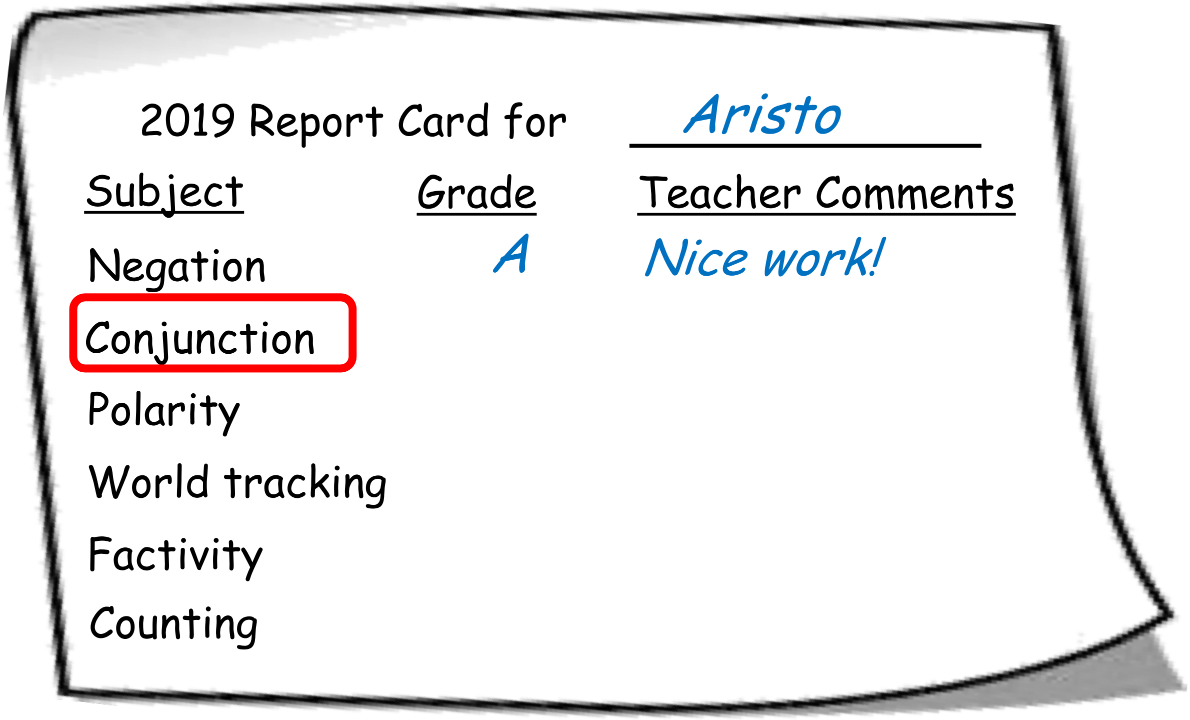
94%

Alan is small. Charlie is big.	Alan is tall. Charlie is tall.	Bob is big. David is small.	Bob is tall. David is short.

Which of the following is ***not*** tall? (A) Alan (B) Bob (C) Charlie (D) David ***[correct]***

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

3. More than Pattern Matching?



2019 Report Card for Aristo

Subject	Grade	Teacher Comments
Negation	A	Nice work!
Conjunction		
Polarity		
World tracking		
Factivity		
Counting		

94%

Synthetic Conjunction Test

Context:

Alan is red.
Alan is big.
Bob is blue.
Bob is small.
Charlie is blue.
Charlie is big.
David is red.
David is small.

Question:

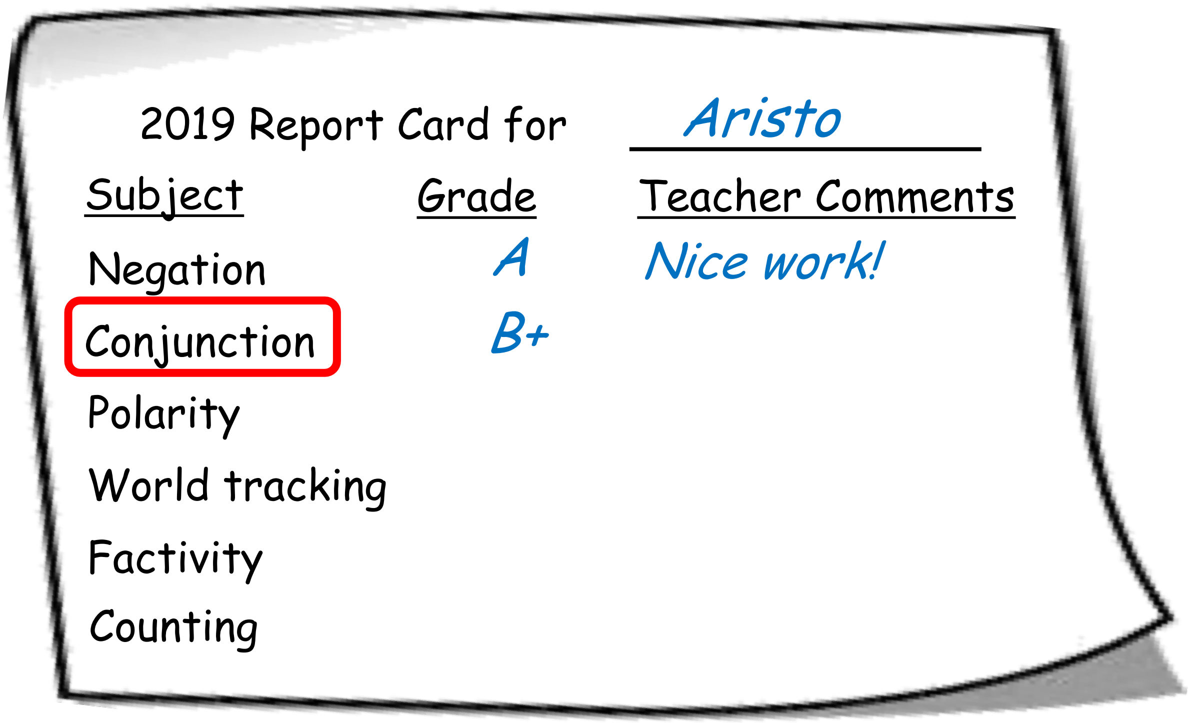
Which of the following is big ***and*** blue? (A) Alan (B) Bob (C) Charlie ***[correct]*** (D) David

		+ 1 negation
1 conjunct:	98%	88.5%
2 conjuncts:	95%	76.5%
3 conjuncts:	94.5%	76%
4 conjuncts:	80%	75%

Alan is red. Alan is big. Alan is light. Alan is old. Alan is tall. Bob is red. Bob is small. Bob is heavy. Bob is old. Bob is tall. Charlie is blue. Charlie is big. Charlie is light. Charlie is old. Charlie is tall. David is red. David is small. David is heavy. David is young. David is tall.

Which of the following is old ***and*** red ***and*** light and big ***and not*** short? (A) Alan (B) Bob (C) Charlie (D) David

3. More than Pattern Matching?



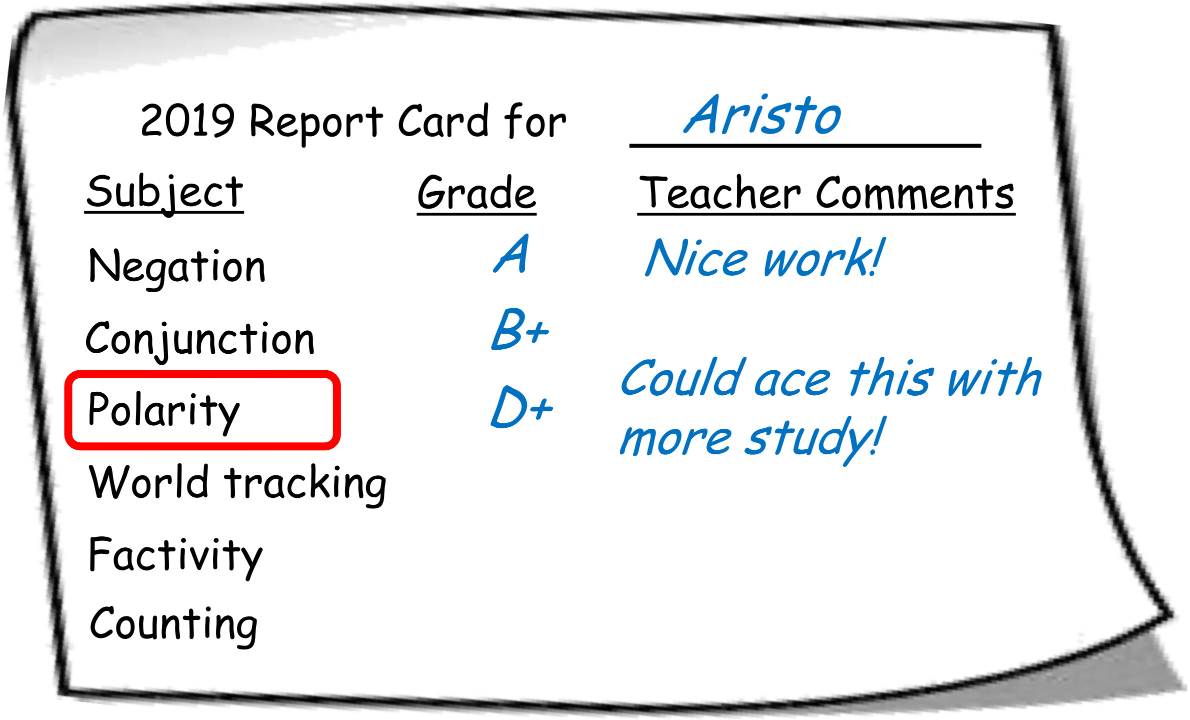
2019 Report Card for *Aristo*

Subject	Grade	Teacher Comments
Negation	A	Nice work!
Conjunction	B+	
Polarity		
World tracking		
Factivity		
Counting		

94%

80% -98%

3. More than Pattern Matching?



2019 Report Card for *Aristo*

Subject	Grade	Teacher Comments
Negation	*A*	*Nice work!*
Conjunction	*B+*	
Polarity	*D+*	*Could ace this with more study!*
World tracking		
Factivity		
Counting		

94%

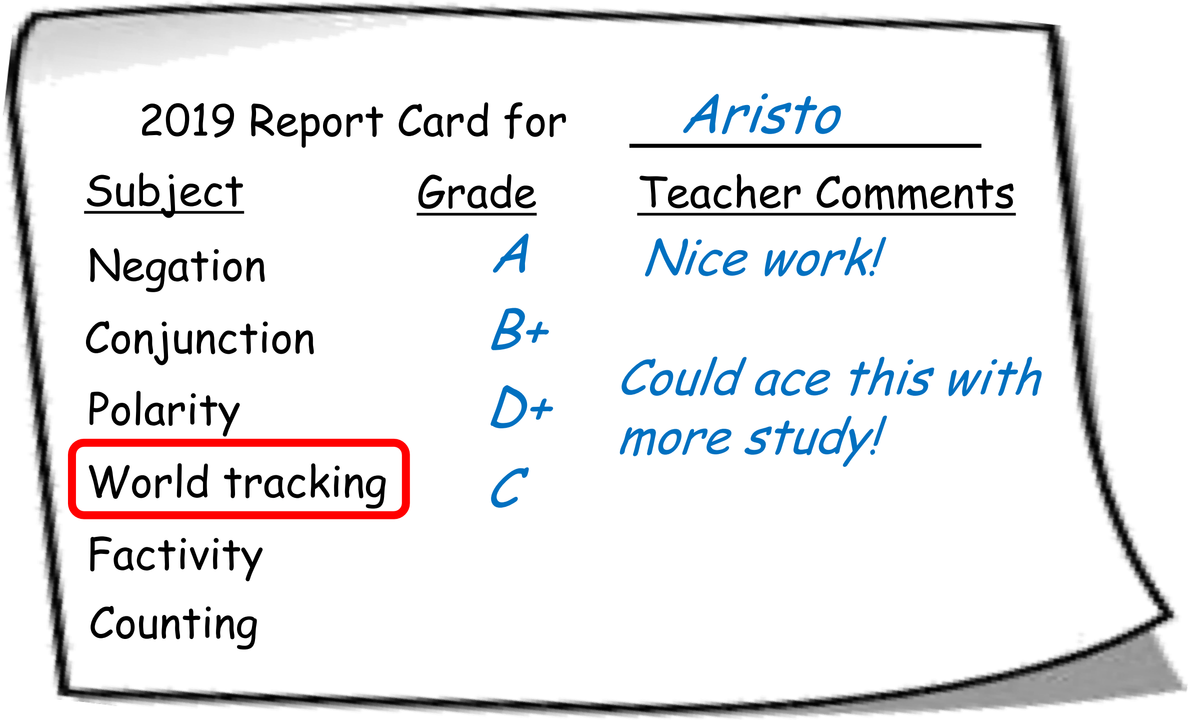
80% -98%

67.1%

Context: For a given medium, sound has a slower speed at lower temperatures.

Question: If Jim turns the thermostat ~~down~~ up in his room while listening to music, what will happen to the speed of the sound waves in the room?
(A) they will speed up [correct] (B) they will slow down ~~***[correct]***~~

3. More than Pattern Matching?



2019 Report Card for Aristo

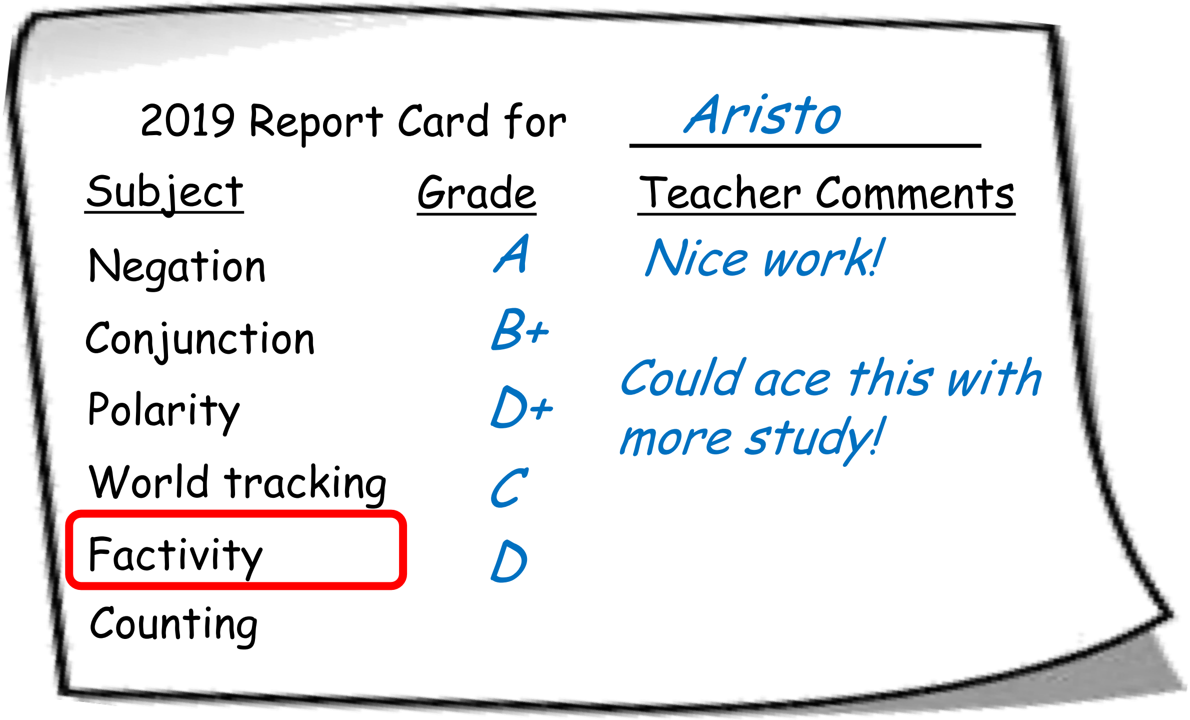
Subject	Grade	Teacher Comments
Negation	A	Nice work!
Conjunction	B+	
Polarity	D+	Could ace this with more study!
World tracking	C	
Factivity		
Counting		

94%
80% -98%
67.1%
72.5%

Context: If someone travels for longer, they will travel further.

Question: John and Rita are going for a run. Rita gets tired and takes a break on the park bench. After twenty minutes in the park, who has run farther?
(A) John ***[correct]*** (B) Rita

3. More than Pattern Matching?



2019 Report Card for *Aristo*

Subject	Grade	Teacher Comments	
Negation	A	Nice work!	94%
Conjunction	B+		80% -98%
Polarity	D+	Could ace this with more study!	67.1%
World tracking	C		72.5%
Factivity	D		66.5%
Counting			

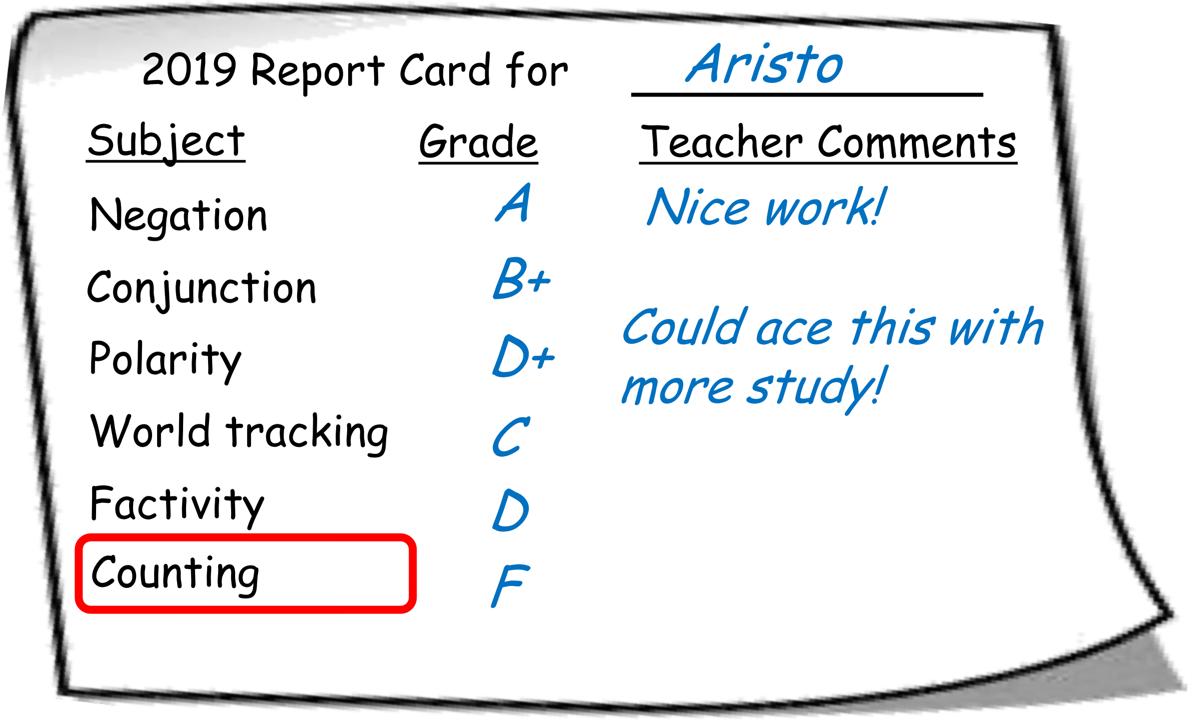
If someone ***regretted*** that a particular thing happened then

(A) that thing might or might not have happened .

(B) that thing didn't happen .

(C) **that thing happened *[correct]***

3. More than Pattern Matching?



2019 Report Card for *Aristo*

Subject	Grade	Teacher Comments	
Negation	A	Nice work!	94%
Conjunction	B+		80% -98%
Polarity	D+	Could ace this with more study!	67.1%
World tracking	C		72.5%
Factivity	D		66.5%
Counting	F		6%

Daniel picked up the football. Daniel dropped the football. Daniel got the milk.

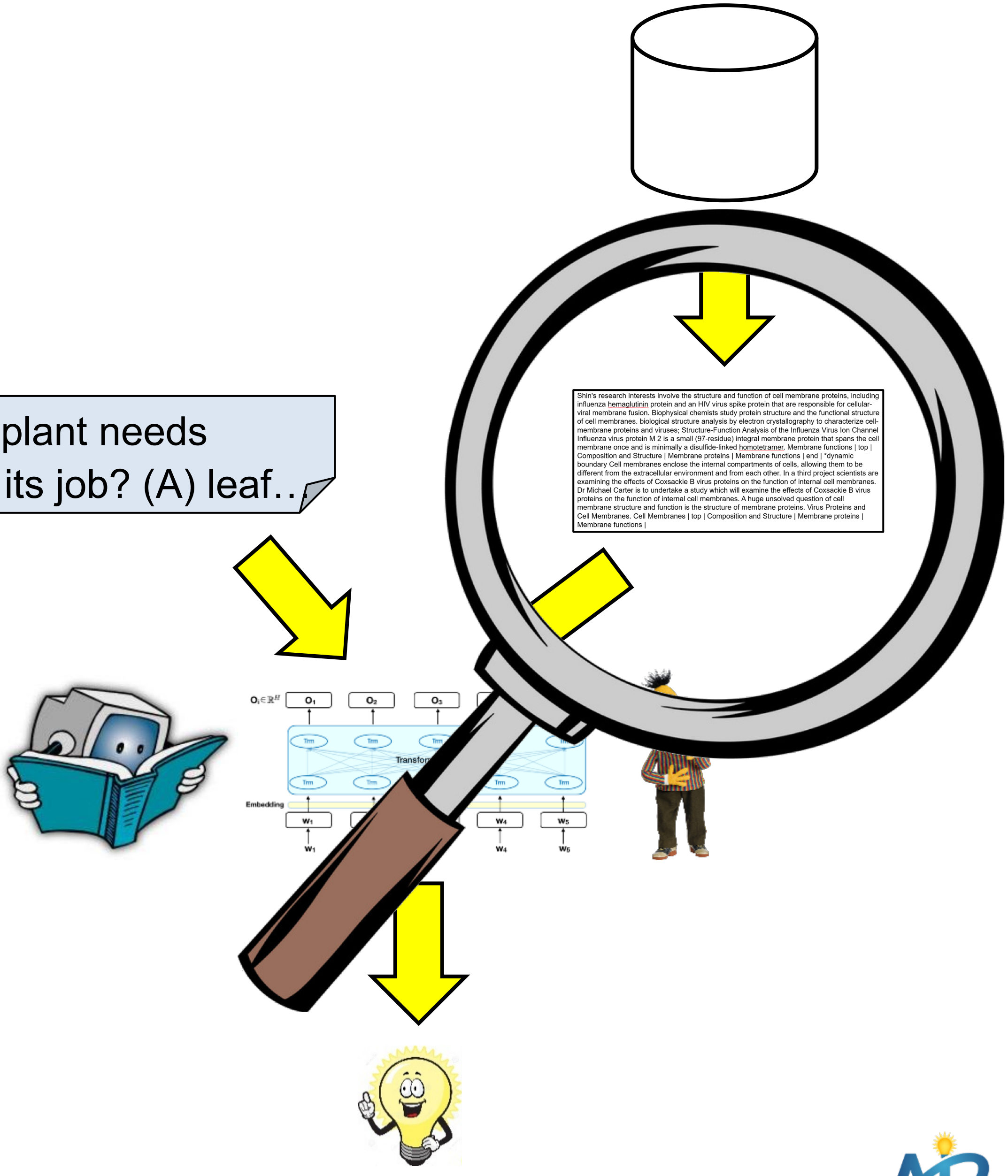
How many objects is Daniel holding? (A) zero **(B) one** (C) two (D) three

Outline

- Introduction
- How does Aristo work?
- What is going on behind the high scores on the exams?
- Where does Aristo fail? ⬅
- What are steps forward?

4. Where is Aristo Failing?

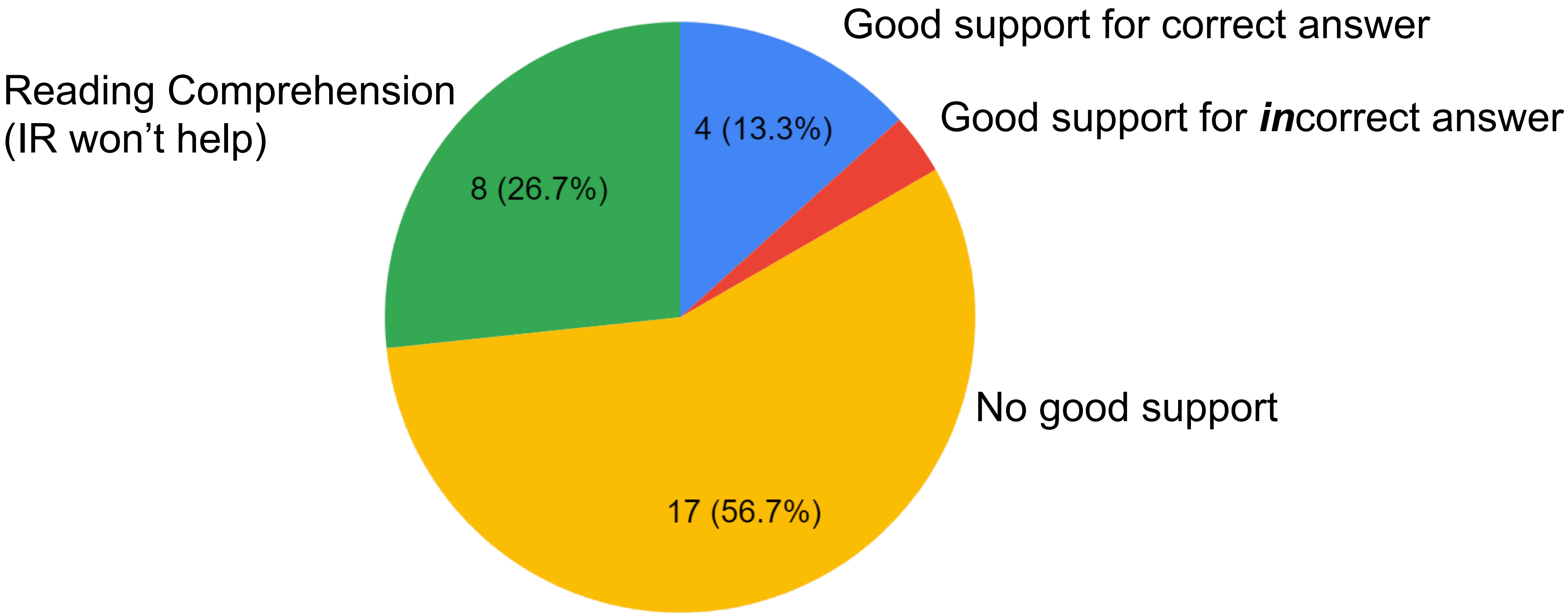
What part of a plant needs sunlight to do its job? (A) leaf..



Shin's research interests involve the structure and function of cell membrane proteins, including influenza hemaglutinin protein and an HIV virus spike protein that are responsible for cellular-viral membrane fusion. Biophysical chemists study protein structure and the functional structure of cell membranes. biological structure analysis by electron crystallography to characterize cell-membrane proteins and viruses; Structure-Function Analysis of the Influenza Virus Ion Channel Influenza virus protein M 2 is a small (97-residue) integral membrane protein that spans the cell membrane once and is minimally a disulfide-linked homotetramer. Membrane functions | top | Composition and Structure | Membrane proteins | Membrane functions | end | *dynamic boundary Cell membranes enclose the internal compartments of cells, allowing them to be different from the extracellular environment and from each other. In a third project scientists are examining the effects of Coxsackie B virus proteins on the function of internal cell membranes. Dr Michael Carter is to undertake a study which will examine the effects of Coxsackie B virus proteins on the function of internal cell membranes. A huge unsolved question of cell membrane structure and function is the structure of membrane proteins. Virus Proteins and Cell Membranes. Cell Membranes | top | Composition and Structure | Membrane proteins | Membrane functions |

4. Where is Aristo failing?

- Case study on 30 failures:



1. Good support for the correct answer (13%)

Which is the best unit to measure distances between Earth and other solar systems in the universe? (A) miles (B) kilometers **(C) light years** **(D) astronomical units**

In general, distances in the solar system are measured in astronomical units.

Distances between Earth and the stars are often measured in terms of light-years. 👍

2. Good support for the ***in***correct answer (3%)

Which of these objects will most likely float in water? (A) glass marble (B) steel ball **(C) hard rubber ball** **(D) table tennis ball**

- *I remember it had like a* ***rubber ball*** *in it, which would maybe* ***float up****…*
- *We played soccer with a giant* ***rubber ball that floated*** *like a balloon.*
- ***Rubber toys floated*** *on the water.*

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

3. No good support for the correct answer (57%)

Although they belong to the same family, an eagle and a pelican are different. What is one difference between them? (A) their preference for eating fish (B) their ability to fly **(C) their method of reproduction** **(D) their method of catching food**

- Need question decomposition

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.

- No good single supporting sentence

ALLEN INSTITUTE
for ARTIFICIAL INTELLIGENCE

3. No good support for the correct answer (57%)

Which characteristic applies to animals in only one of these taxonomic groups: reptiles, mammals, birds, amphibians, or fishes? **(A) have hair** **(B) lay eggs** (C) have webbed feet (D) breathe with gills

- Boolean reasoning

Which geologic structure will most likely take the longest time to form? (A) a fault (B) a sinkhole **(C) a river meander** **(D) a mountain range**

- Cross-option comparative

4. Reading Comprehension (27%)

- Story (experimental method)

A student wants to determine the effect of garlic on the growth of a fungus species. Several samples of fungus cultures are grown in the same amount of agar and light. Each sample is given a different amount of garlic. What is the independent variable in this investigation? (A) amount of agar (B) amount of light **(C) amount of garlic** **(D) amount of growth**

- Meta/sentiment

Which statement is an opinion? (A) Many plants are green. **(B) Many plants are beautiful.** **(C) Plants require sunlight.** (D) Plants can grow in different places.

Math Reasoning

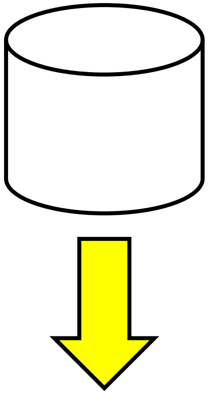
About how long does it take for the Moon to complete one revolution around Earth? (A) 7 days **(B) 30 days** (C) 90 days (**D) 365 days**

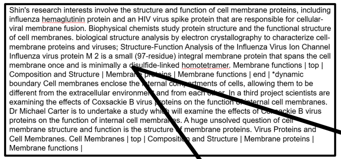
- *Because it takes the moon about* ***27.3 days*** *to complete one orbit around the Earth, the moon moves a little bit further around the Earth each day.*
- *It takes* ***27.3 days*** *for the moon to complete one revolution around the earth.*
- *The moon completes one revolution of the Earth in about* ***29.5 days.***
- *The Moon completes one revolution around the Earth in* ***27.32166 days.***

Outline

- Introduction
- How does Aristo work?
- What is going on behind the high scores on the exams?
- Where does Aristo fail?
- What are steps forward?

1. Question Decomposition





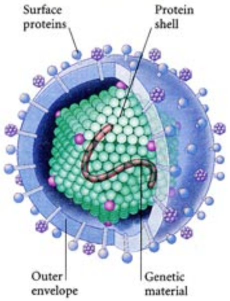
Shin's research interests involve the structure and function of cell membrane proteins, including influenza hemaglutinin protein and an HIV virus spike protein that are responsible for cellular-viral membrane fusion. Biophysical chemists study protein structure and the functional structure of cell membranes. biological structure analysis by electron crystallography to characterize cell-membrane proteins and viruses; Structure-Function Analysis of the Influenza Virus Ion Channel Influenza virus protein M 2 is a small (97-residue) integral membrane protein that spans the cell membrane once and is minimally a disulfide-linked homotetramer. Membrane functions | top | Composition and Structure | Membrane proteins | Membrane functions | end | *dynamic boundary Cell membranes enclose the internal compartments of cells, allowing them to be different from the extracellular environment and from each other. In a third project scientists are examining the effects of Coxsackie B virus proteins on the function of internal cell membranes. Dr Michael Carter is to undertake a study which will examine the effects of Coxsackie B virus proteins on the function of internal cell membranes. A huge unsolved question of cell membrane structure and function is the structure of membrane proteins. Virus Proteins and Cell Membranes. Cell Membranes | top | Composition and Structure | Membrane proteins | Membrane functions |

What virus structure is similar in function to a cell membrane?
(A) protein shell (B) internal protein...

structure-function of membrane proteins. membrane protein structure and function; Structure and function of membrane proteins; Shin's research interests involve the structure and function of cell membrane proteins, including influenza hemaglutinin protein and an HIV virus spike protein that are responsible for cellular-viral membrane fusion. biological structure analysis by electron crystallography to characterize cell-membrane proteins and viruses; Structure-Function Analysis of the Influenza Virus Ion Channel Influenza virus protein M 2 is a small (97-residue) integral membrane protein that spans the cell membrane once and is minimally a disulfide-linked homotetramer. Biophysical chemists study protein structure and the functional structure of cell membranes. A huge unsolved question of cell membrane structure and function is the structure of membrane proteins. Virus Proteins and Cell Membranes. Cell Membranes | top | Composition and Structure | Membrane proteins | Membrane functions |

1. Question Decomposition

What virus structure is similar in function to a cell membrane?
(A) protein shell (B) internal protein...



What is the function of a cell membrane?

Surrounds and protects, gives structure, regulates material, ….

What part of the virus surrounds and protects it?

Protein shell, protein layer, …

- GapQA *(EMNLP'19)*
- New dataset coming

2. Multihop Reasoning

Which conducts electricity? (A) suit of armor (B) cotton candy

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

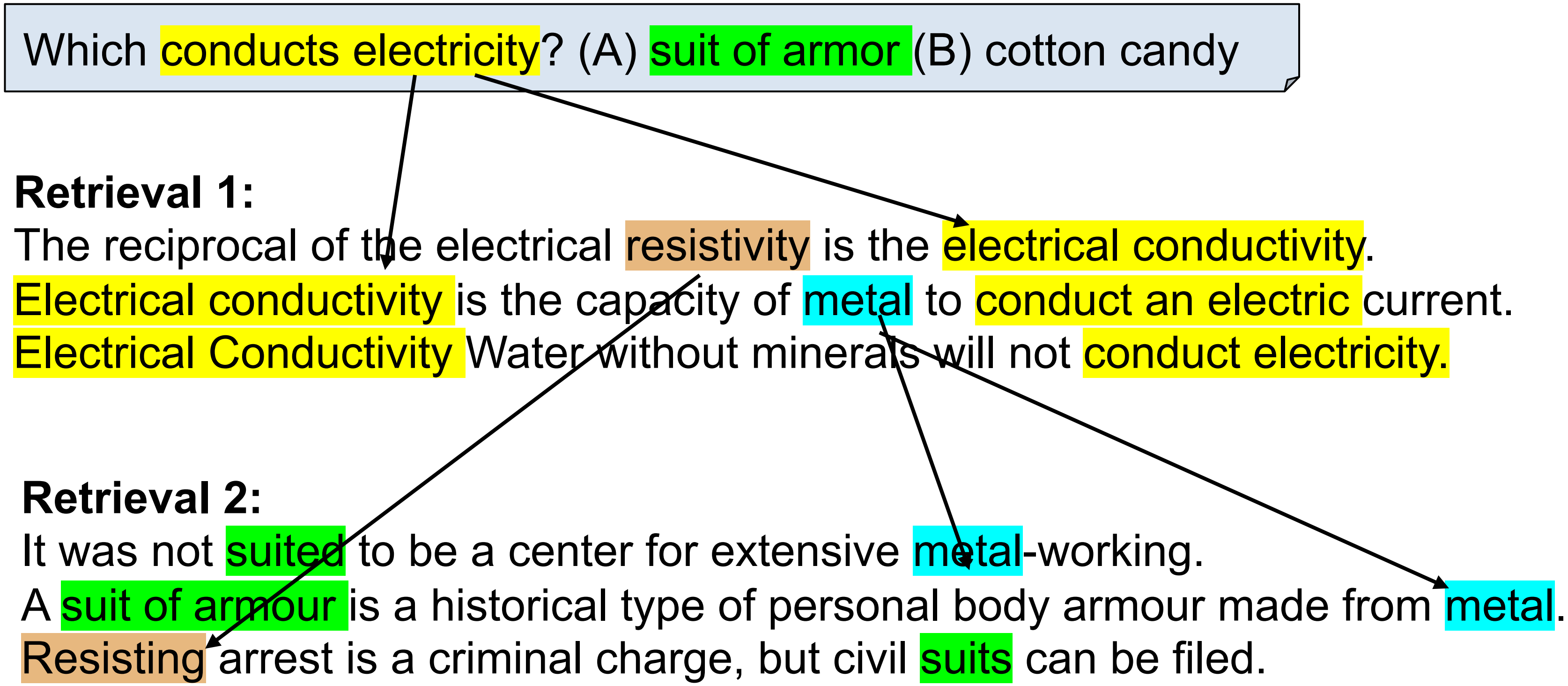
2. Multihop Reasoning

Which conducts electricity? (A) suit of armor (B) cotton candy

Retrieval 1:

The reciprocal of the electrical resistivity is the electrical conductivity.
Electrical conductivity is the capacity of metal to conduct an electric current.
Electrical Conductivity Water without minerals will not conduct electricity.

2. Multihop Reasoning



Which conducts electricity? (A) suit of armor (B) cotton candy

Retrieval 1:
The reciprocal of the electrical resistivity is the electrical conductivity.
Electrical conductivity is the capacity of metal to conduct an electric current.
Electrical Conductivity Water without minerals will not conduct electricity.

Retrieval 2:
It was not suited to be a center for extensive metal-working.
A suit of armour is a historical type of personal body armour made from metal.
Resisting arrest is a criminal charge, but civil suits can be filed.

Form Chains:

"suit of armor…made from metal" AND "…metal conduct electrical current" ☺
=> "suit of armor conducts electricity"

"Resisting arrest…suits can be filed" AND "reciprocal of resistivity is conductivity" ☹
=> "suit of armor conducts electricity"

Train system to recognize good chains

AI2 ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

3. Modeling World States

Photosynthesis

Roots absorb water from the soil.

The water flows to the leaf.

Light and CO2 enter leaf.

Light, water, CO2 form sugar.

3. Modeling World States

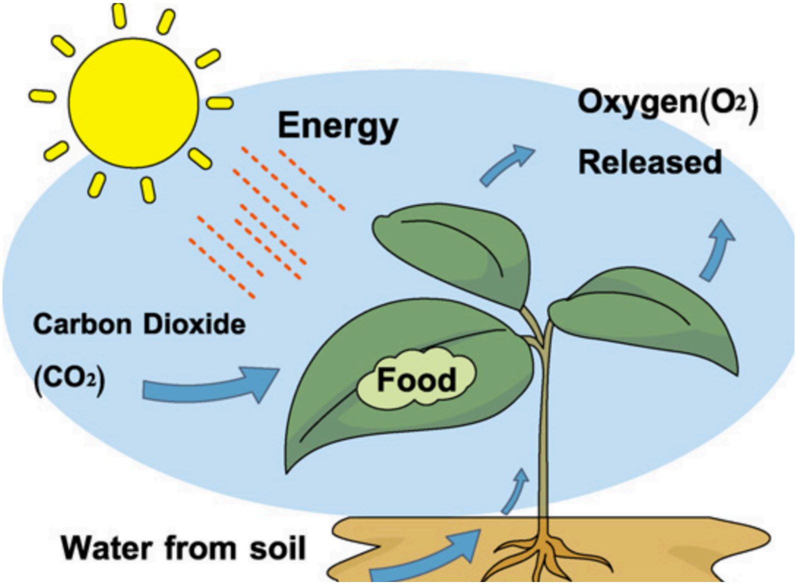
Photosynthesis

Roots absorb water from the soil.

The water flows to the leaf.

Light and CO2 enter leaf.

Light, water, CO2 form sugar.



Where is the sugar created? **Light, water, CO2** [BiDAF]

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

3. Modeling World States

Paragraph

s1	Roots absorb water from soil.
s2	The water flows to the leaf.
s3	Light and CO2 enter leaf.
s4	Water, light, CO2 form sugar.

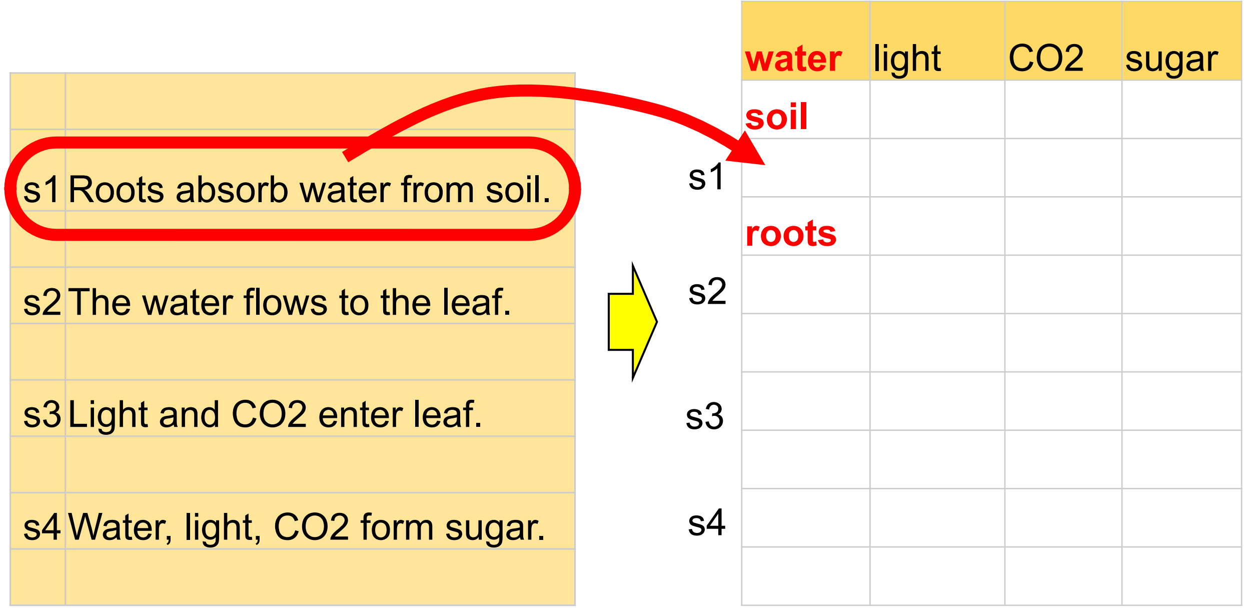
State changes: π

	water	light	CO2	sugar
s1				
s2				
s3				
s4				

3. Modeling World States

Paragraph

State changes: *π*



s1	Roots absorb water from soil.
s2	The water flows to the leaf.
s3	Light and CO2 enter leaf.
s4	Water, light, CO2 form sugar.

	water	light	CO2	sugar
	soil			
s1				
	roots			
s2				
s3				
s4				

3. Modeling World States

Paragraph

s1	Roots absorb water from soil.
s2	The water flows to the leaf.
s3	Light and CO2 enter leaf.
s4	Water, light, CO2 form sugar.

State changes: *π*

	water	light	CO2	sugar
	soil			
s1				
	roots			
s2				
s3				
s4				

ALLEN INSTITUTE
for ARTIFICIAL INTELLIGENCE

3. Modeling World States

Paragraph

s1	Roots absorb water from soil.
s2	The water flows to the leaf.
s3	Light and CO2 enter leaf.
s4	Water, light, CO2 form sugar.

State changes: π

	water	light	CO2	sugar
	soil			
s1				
	roots			
s2				
	leaf			
s3				
s4				

ALLEN INSTITUTE
for ARTIFICIAL INTELLIGENCE

3. Modeling World States

Paragraph

s1	Roots absorb water from soil.
s2	The water flows to the leaf.
s3	Light and CO2 enter leaf.
s4	Water, light, CO2 form sugar.

State changes: π

	water	light	CO2	sugar
	soil			
s1				
	roots			
s2				
	leaf			
s3				
	leaf			
s4				

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

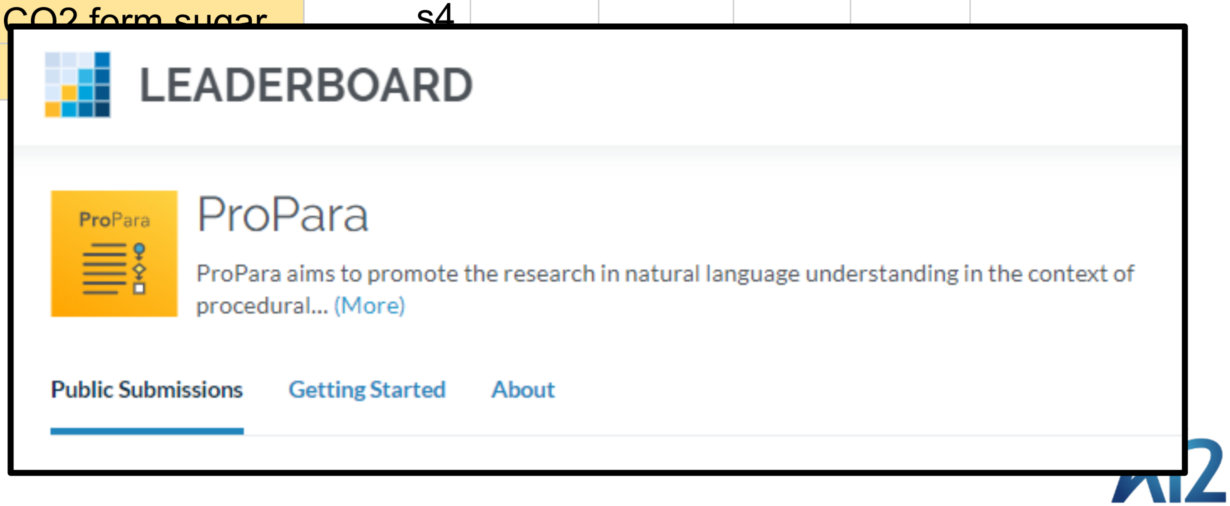
3. Modeling World States

Paragraph

s1	Roots absorb water from soil.
s2	The water flows to the leaf.
s3	Light and CO2 enter leaf.
s4	Water, light, CO2 form sugar

State changes: π

	water	light	CO2	sugar
	soil	sun	?	-
s1				
	roots	sun	?	-
s2				
	leaf	sun	?	-
s3				
	leaf	leaf	leaf	-
s4				



LEADERBOARD

ProPara

ProPara aims to promote the research in natural language understanding in the context of procedural... (More)

4. Explanation and Instruction

Can you pick up a penny with a magnet?

Yes

Why?

Because

- *pennies are made of metal*
- *metals are magnetic*

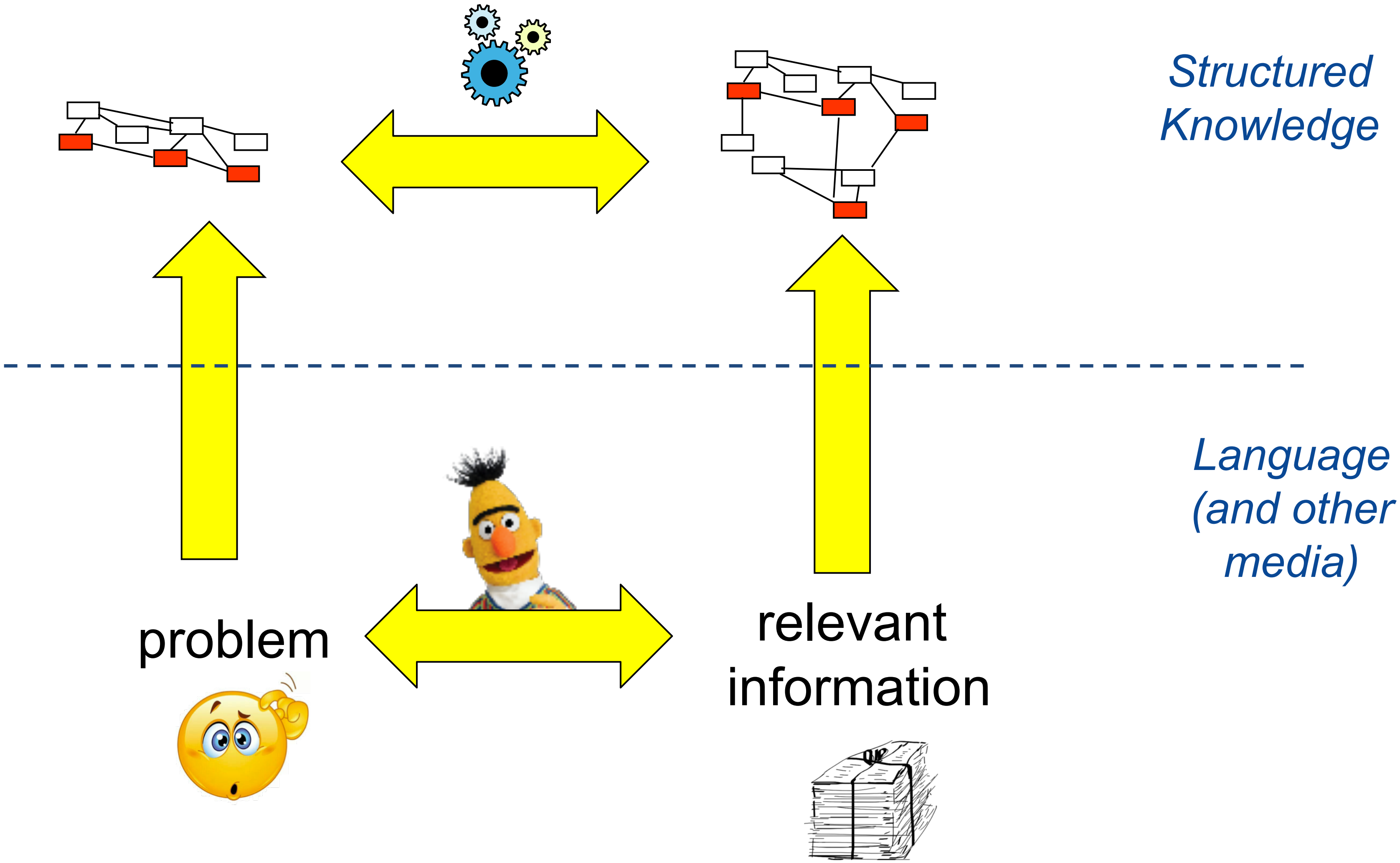
Actually:
Not all metals are magnetic.
Copper is not magnetic.

Try again!

No – because:

- *pennies are made of copper*
- *copper is not magnetic*

A Question for the field of Knowledge Capture



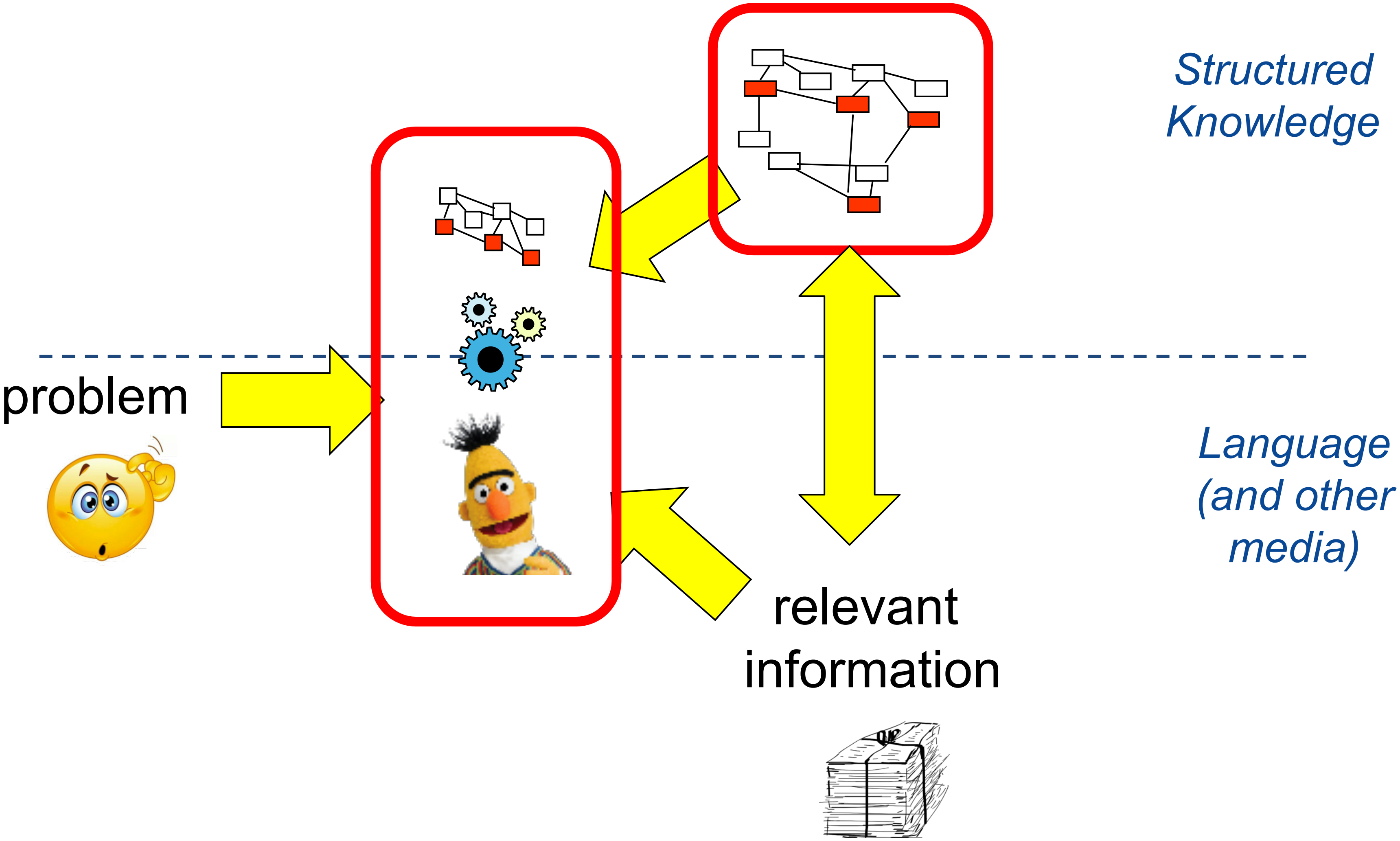
Structured Knowledge

Language (and other media)

problem

relevant information

A Question for the field of Knowledge Capture



Structured Knowledge

problem

relevant information

Language (and other media)

Summary

ARISTO

- Surprising success!
  - LMs: Structure not essential for many tasks
  - >> “just pattern matching”
- BUT:
  - falls short with numerous types of questions
  - many other AI aspects missing

What do we need going forward?

- Structured reasoning and knowledge capture *but* with more language-like representations

Thank you!

AI2 ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE