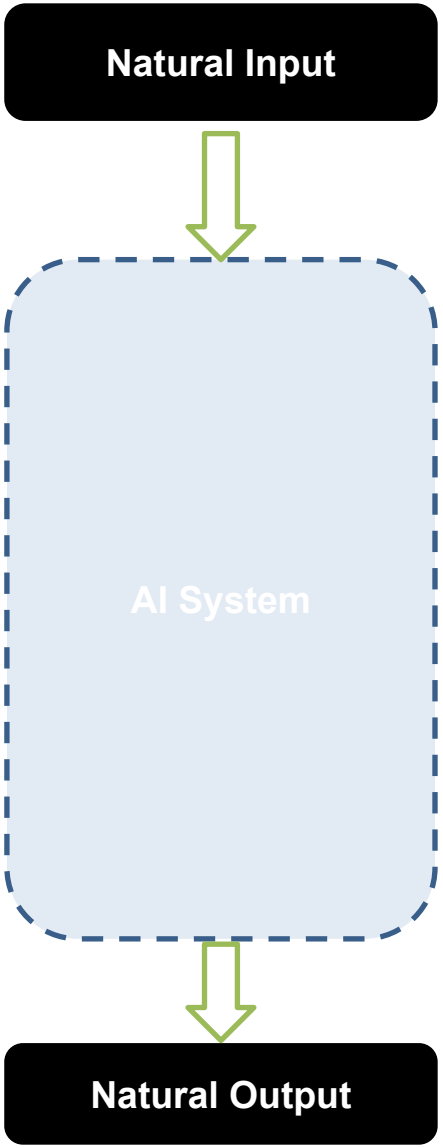
Reasoning-driven Question Answering

Daniel Khashabi, Dan Roth
(Cognitive Computation Group, UPenn)

Tushar Khot, Ashish Sabharwal
(Allen Institute for Artificial Intelligence)

QA protocol and the magic box



“A student leaves a bucket of water outside on a warm, sunny day. The water disappeared after a few hours. What phenomena describes this?”

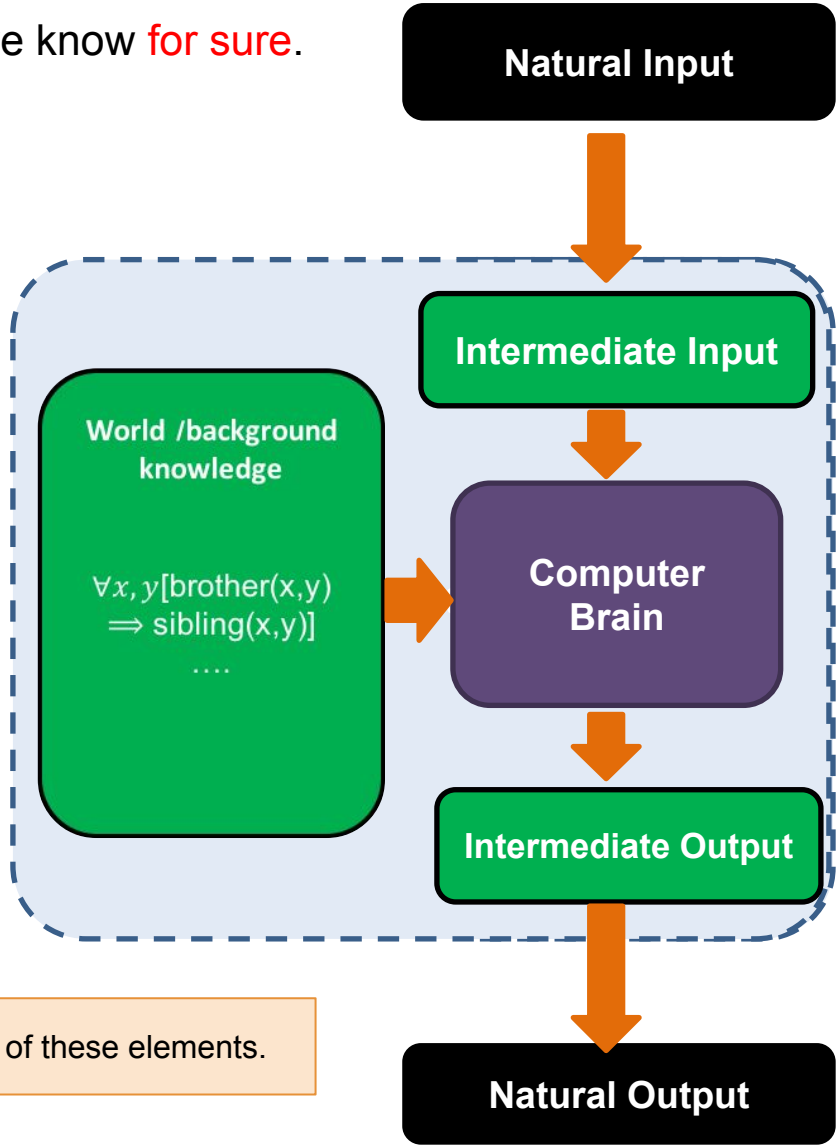
“Evaporation”

What we know for sure

But there are certain things that we know for sure.

A "good" solution has to have:

- knowledge
- knowledge representation
- "easy" way of accessing the knowledge
- a decision making mechanism



"A student leaves a bucket of water outside on a warm, sunny day. The water disappeared after a few hours. What phenomena describes this?"

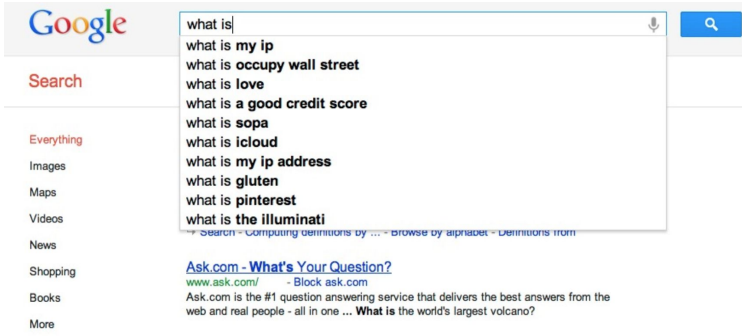
"Evaporation"

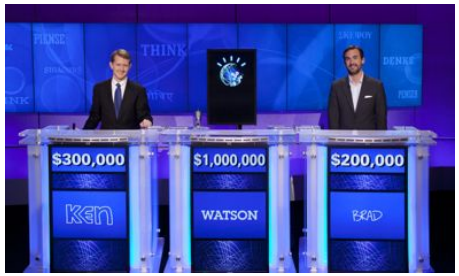
This does not entail isolation/independence of these elements.

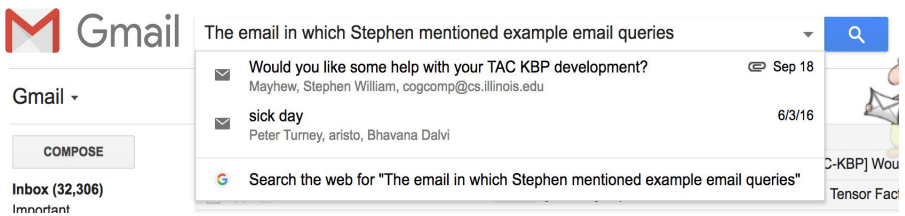
Penn
University of Pennsylvania

QA is everywhere

- One of the oldest problems in AI
- Remarkable features of QA







QA systems are still far from exhibiting human-like intelligence, even in relatively simple ways (vs. human-level)

General Problem Solver

(Simon&Newell, 1956)

Goal: Program for proving theorems!

Necessity: Representation with symbols!

Hypothesis (physical symbol system hypothesis):
"A physical symbol system has the necessary and sufficient means for general intelligent action."

Reasoning: Problem solving as Search!

Programs with Commonsense

(John McCarthy, 1959)

Formalize world in **logical** form!

Example:
"My desk is at home" → at(I, desk)
"Desk is at home" → at(desk, home)

Hypothesis: Commonsense knowledge can be formalized with logic.

Do **reasoning** on formal premises!

Example Contd.:
$\forall x \forall y \forall z$ at(x,y), at(y,z)→ at(x, z)
∴ at(I, home)

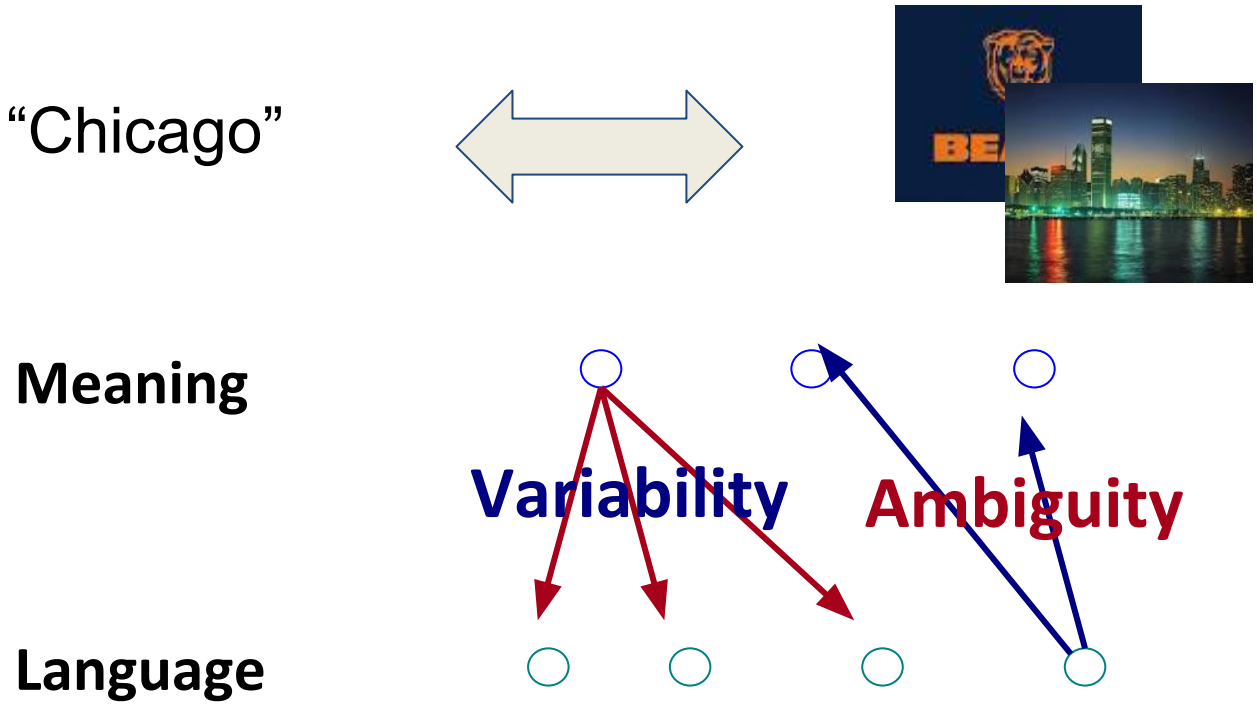
Hypothesis: Commonsense problems are solved by logical reasoning

Penn
University of Pennsylvania

What they missed ….

- The difficulty of mapping from nature (including natural language) to symbols

One cannot simply map natural language to a representation that gives rise to reasoning

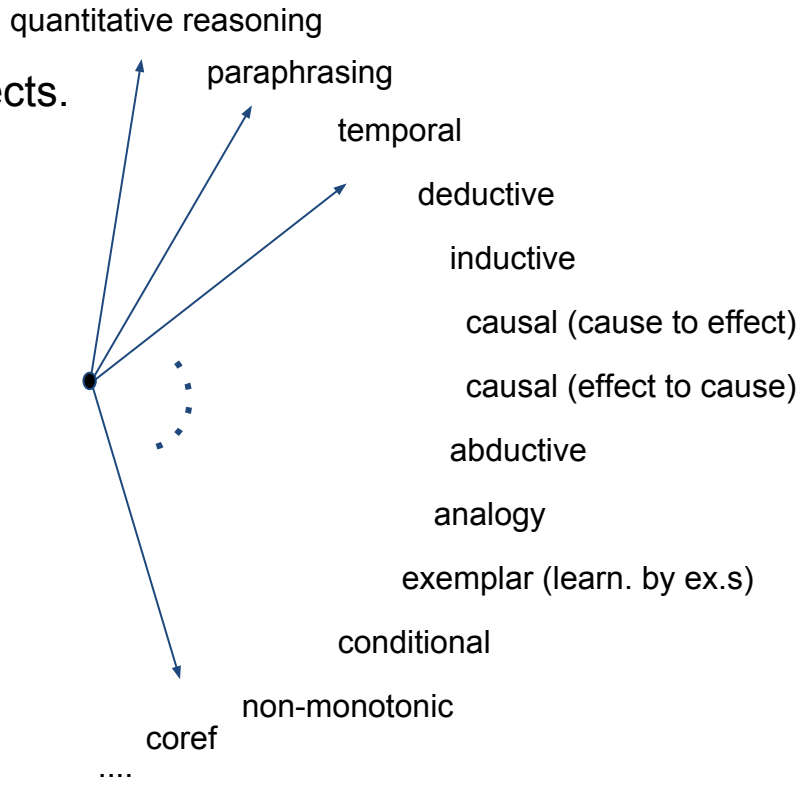


What they missed ….

- Reasoning is often studied in a very narrow sense.

Reasoning has many (infinite?) forms.

- One can think of it as a *n* dimensional space
- Examples typically span multiple reasoning aspects.



How do you define “reasoning”?

It’s a convoluted subject and hard to define.

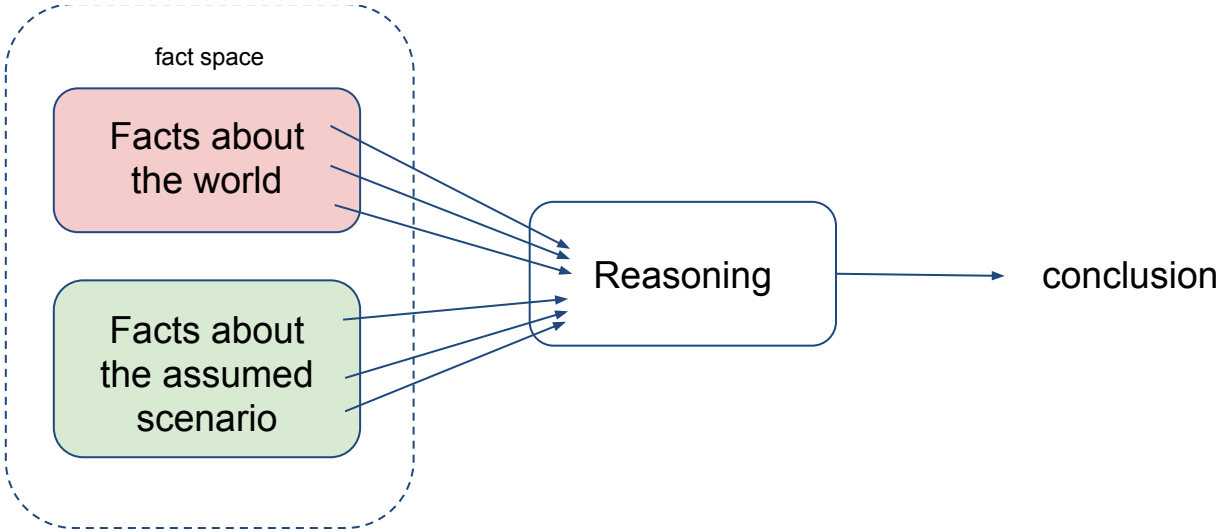
reasoning in British
(ˈriːzənɪŋ)

Word Frequency ●●●●●

noun

1. the act or process of drawing conclusions from facts, evidence, etc
2. the arguments, proofs, etc, so adduced

Collins English Dictionary. Copyright © HarperCollins Publishers

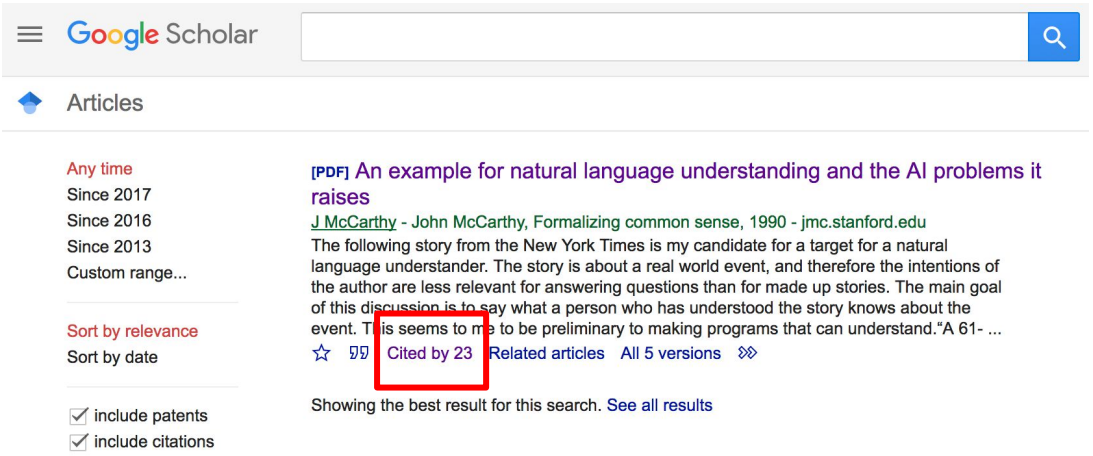


An Example for NLU …

AN EXAMPLE FOR NATURAL LANGUAGE UNDERSTANDING AND THE AI PROBLEMS IT RAISES

John McCarthy
Computer Science Department
Stanford University
Stanford, CA 94305
jmc@cs.stanford.edu
http://www-formal.stanford.edu/jmc/

1976



Penn
University of Pennsylvania

An Example for NLU …

(John McCarthy, 1976)

Taken from NYT (natural text)

A 61-year old furniture salesman was pushed down the shaft of a freight elevator yesterday in his downtown Brooklyn store by two robbers while a third attempted to crush him with the elevator car because they were dissatisfied with the $1,200 they had forced him to give them.
The buffer springs at the bottom of the shaft prevented the car from crushing the salesman, John J. Hug, after he was pushed from the first floor to the basement. The car stopped about 12 inches above him as he flattened himself at the bottom of the pit.
… (McCarthy, 1990, p. 70)

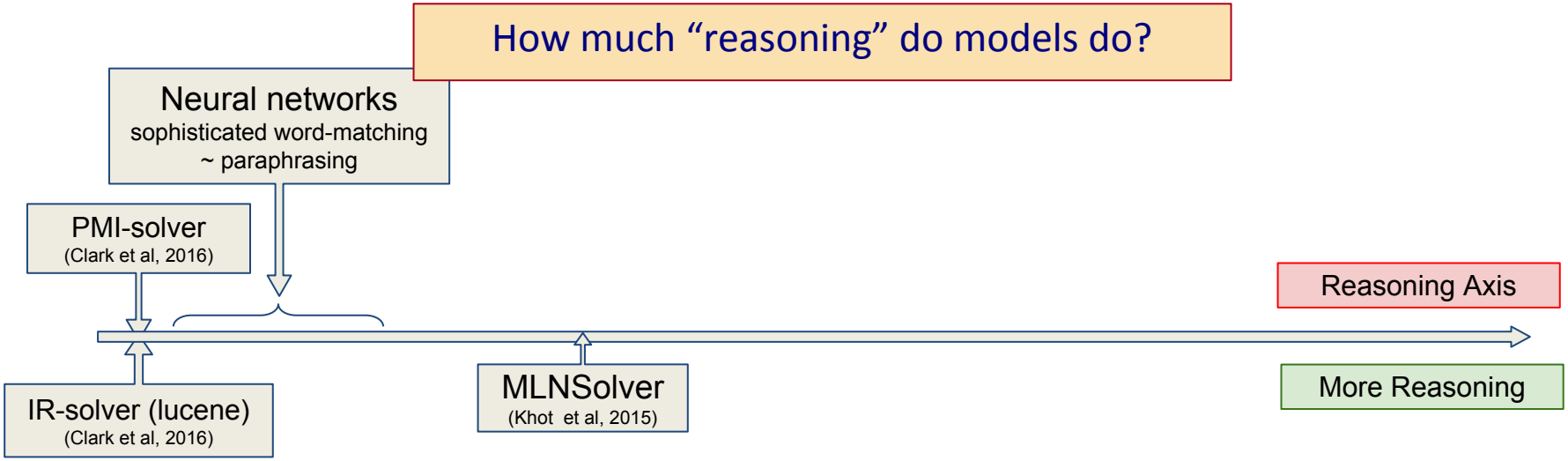
- Predictions: Who had the money at the end?
- Yes/No question: Did Mr. Hug want to be crushed?
- What if question: What would have happened if Mr. Hug had not flattened himself at the bottom of the pit?

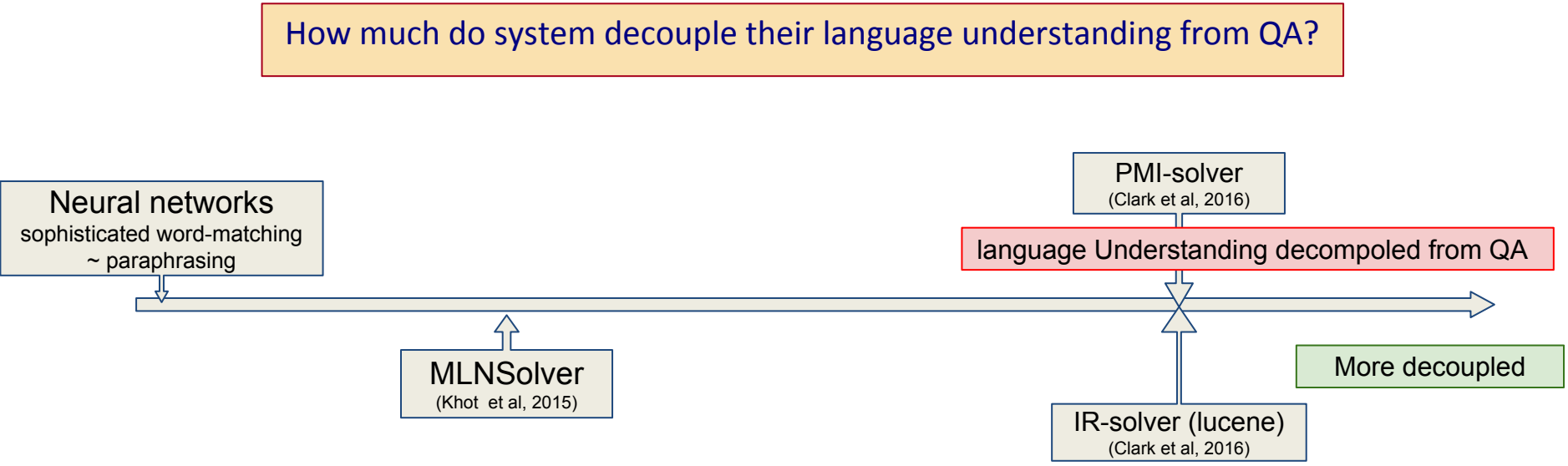
…

Not mentioned directly. It is implied.

Requires common sense + changing the knowledge according to the alternative scenario

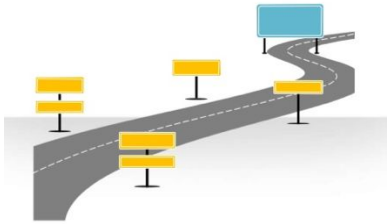
Where are we?





The Roadmap

- **Motivation**
  - Reasoning: Past and now

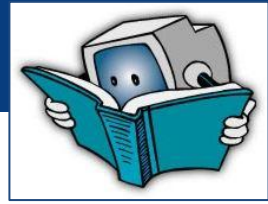


- **TableILP:** science QA with tables a knowledge
  1. Motivating example
  2. Tables as Knowledge
  3. Tables + ILP = TableILP solver

- **SemanticILP:** reasoning with layers of representation
  1. Motivating example
  2. Text + off-the-shelf annotators as knowledge
  3. Layers of semantic representations + ILP = SemanticILP solver

- **Summary**

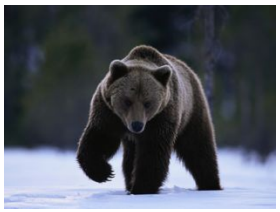
Standardized Tests as an AI Challenge



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



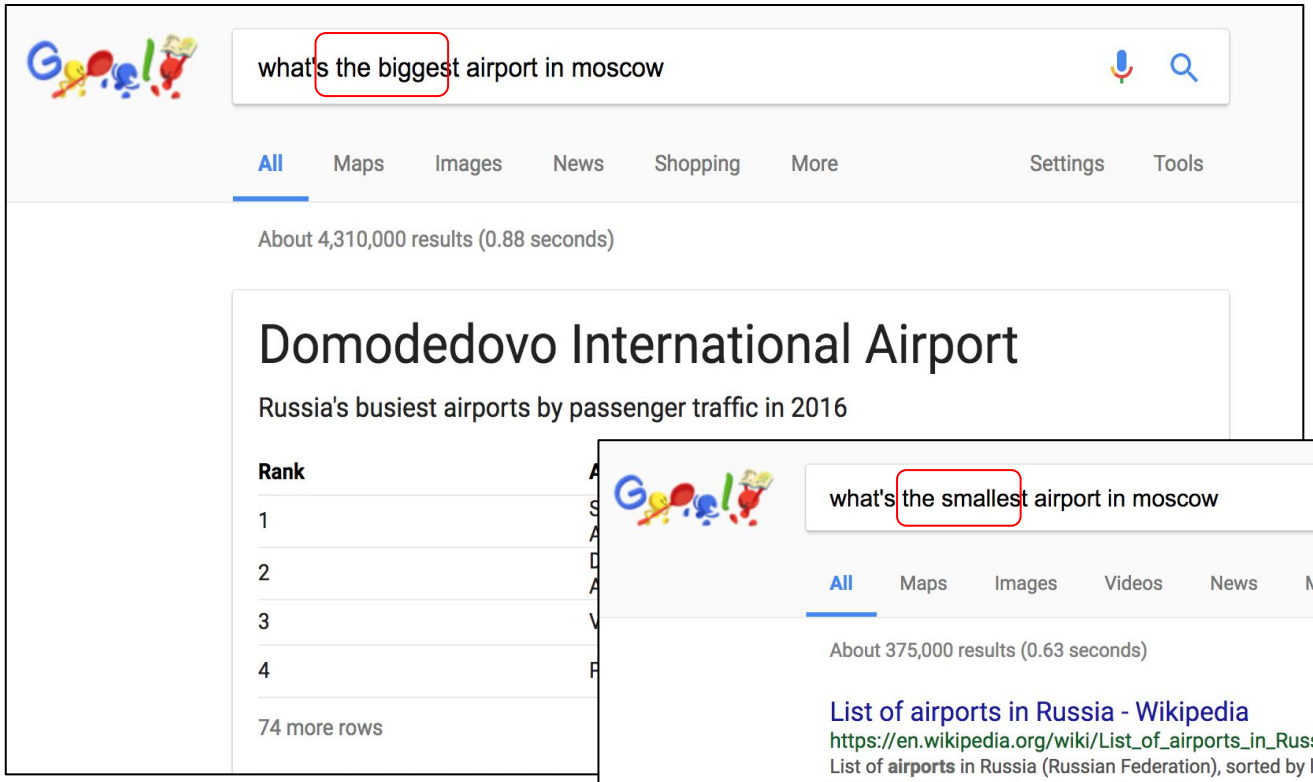
New Zealand shortest night

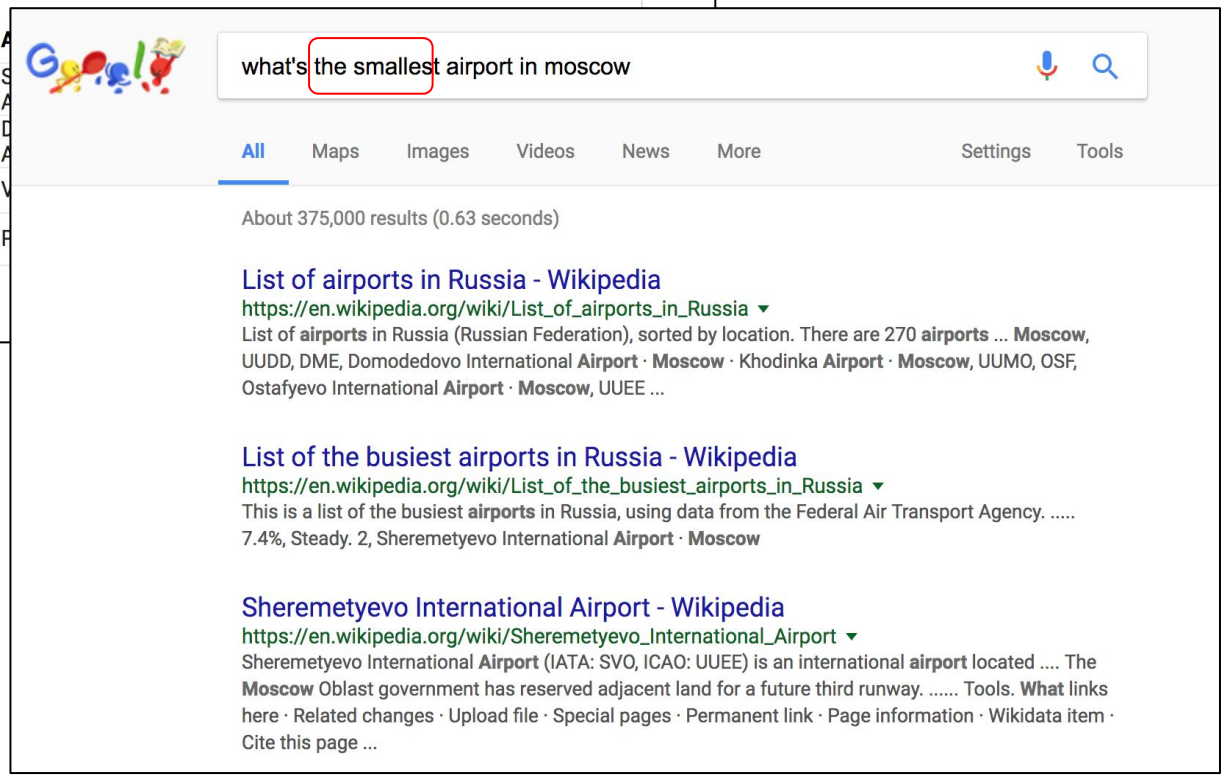
In ~~New York State~~, the ~~longest~~ period of ~~daylight~~ occurs during which month?

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

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

Google is missing this behavior





Semi-Structured Inference

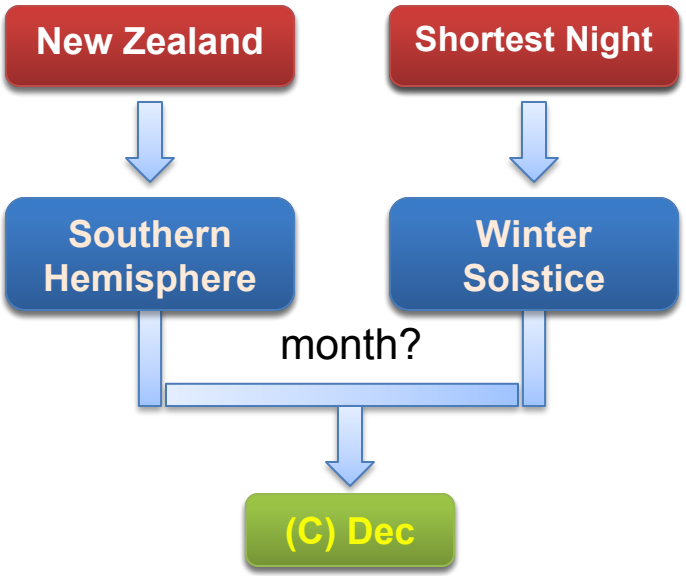
New Zealand shortest night

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

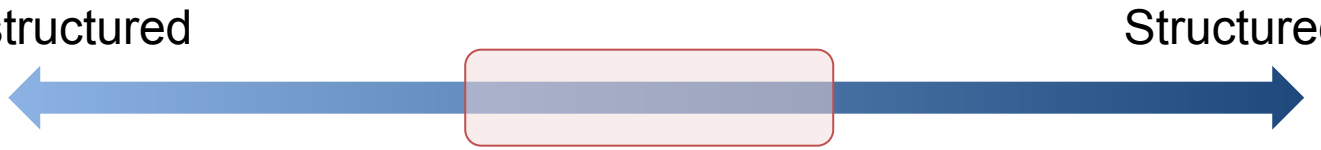
✓ principled approach, explainable answers
✓ robust to variations

How can we achieve this?



Penn University of Pennsylvania

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

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

TableILP: Main Idea

Search for the best **Support Graph** connecting the Question to an Answer through Tables (i.e best explanation)

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

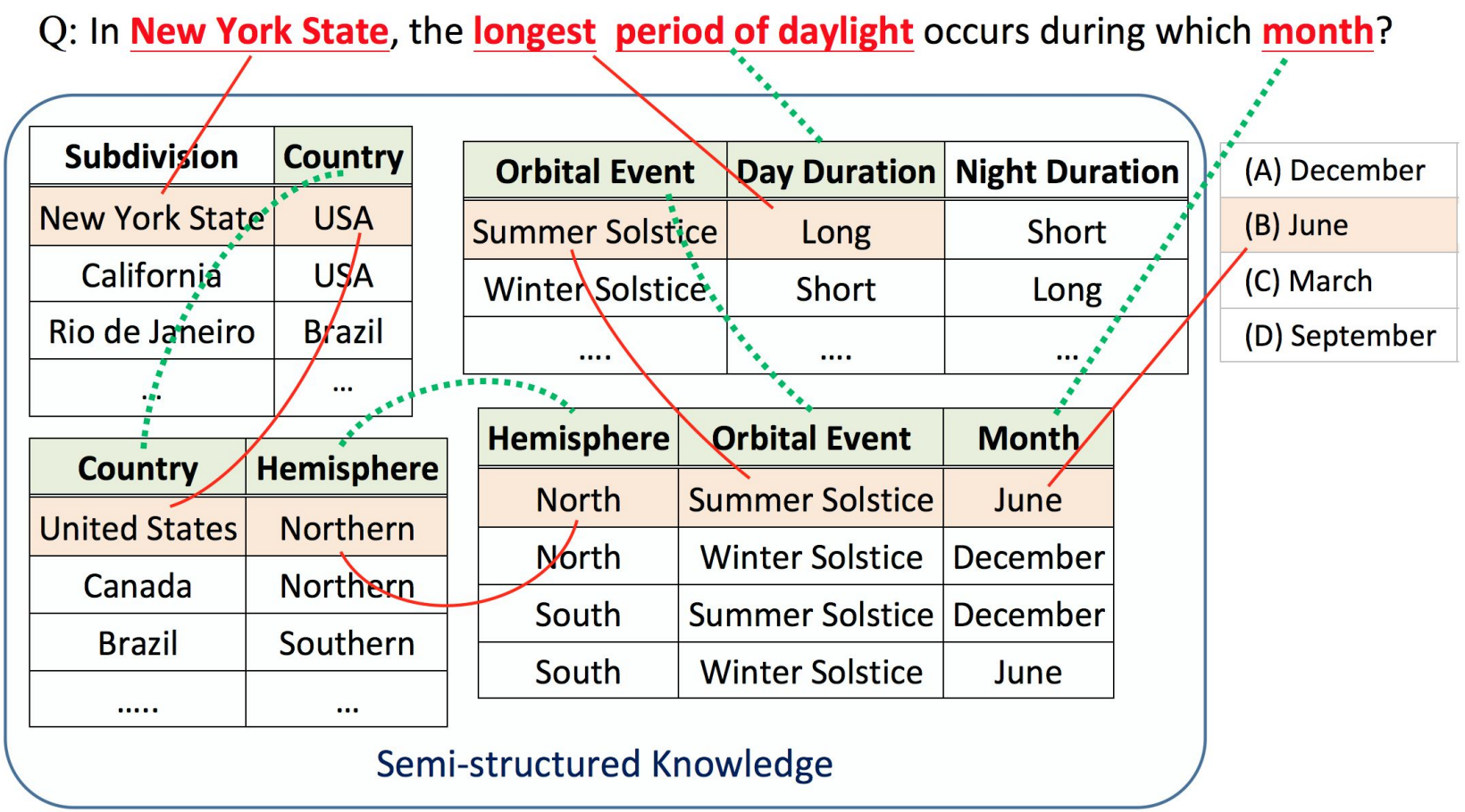
Semi-structured Knowledge

(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!

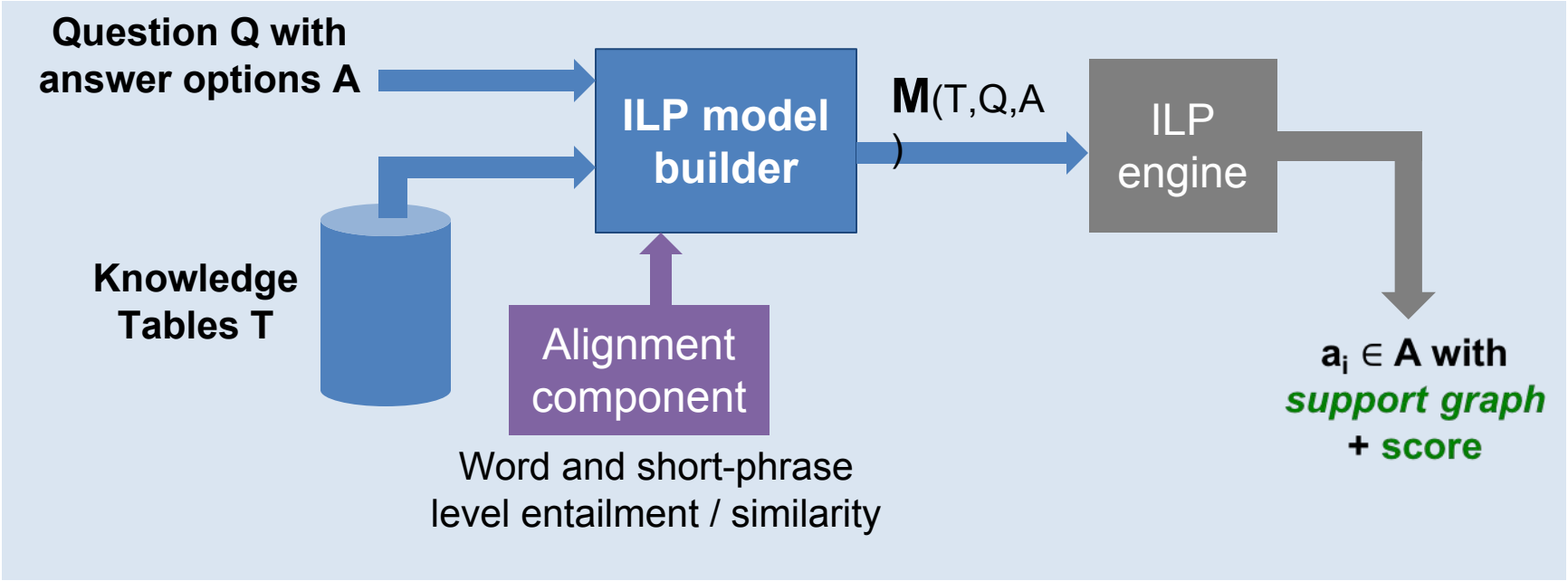


Penn University of Pennsylvania

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) →

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

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

Optimization using Integer Linear Prog. formalism

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

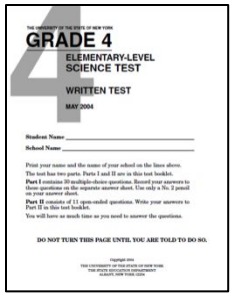
- Reward active units, high lexical match links, column header match, …
- Penalize spurious overuse of frequently occurring terms

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 _____

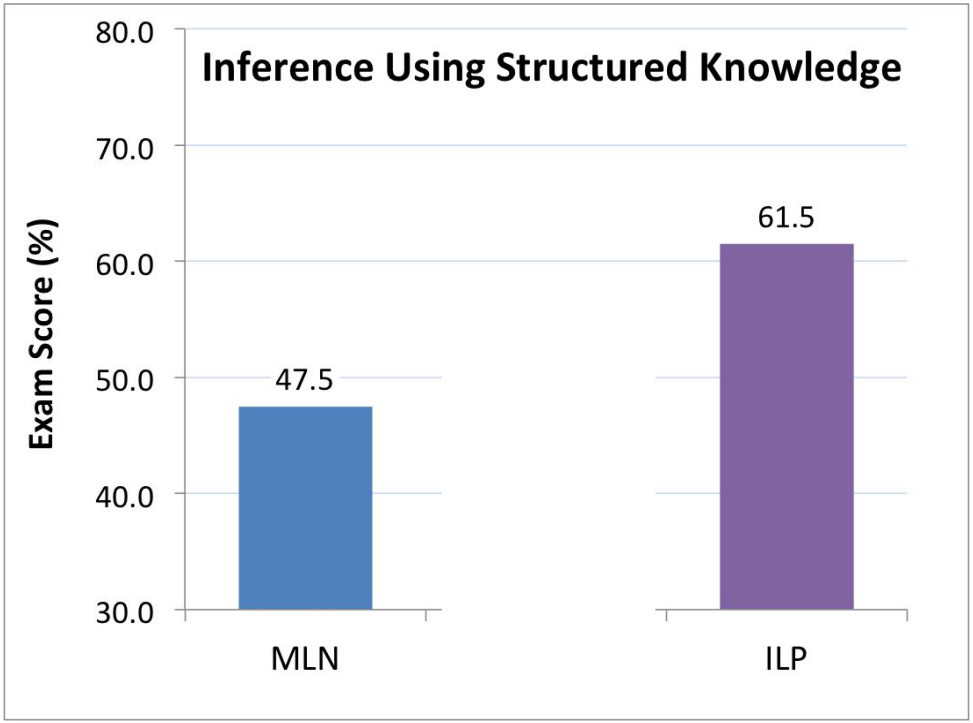
DO NOT TURN THIS PAGE UNTIL YOU ARE TOLD TO DO SO.

Available at allenai.org

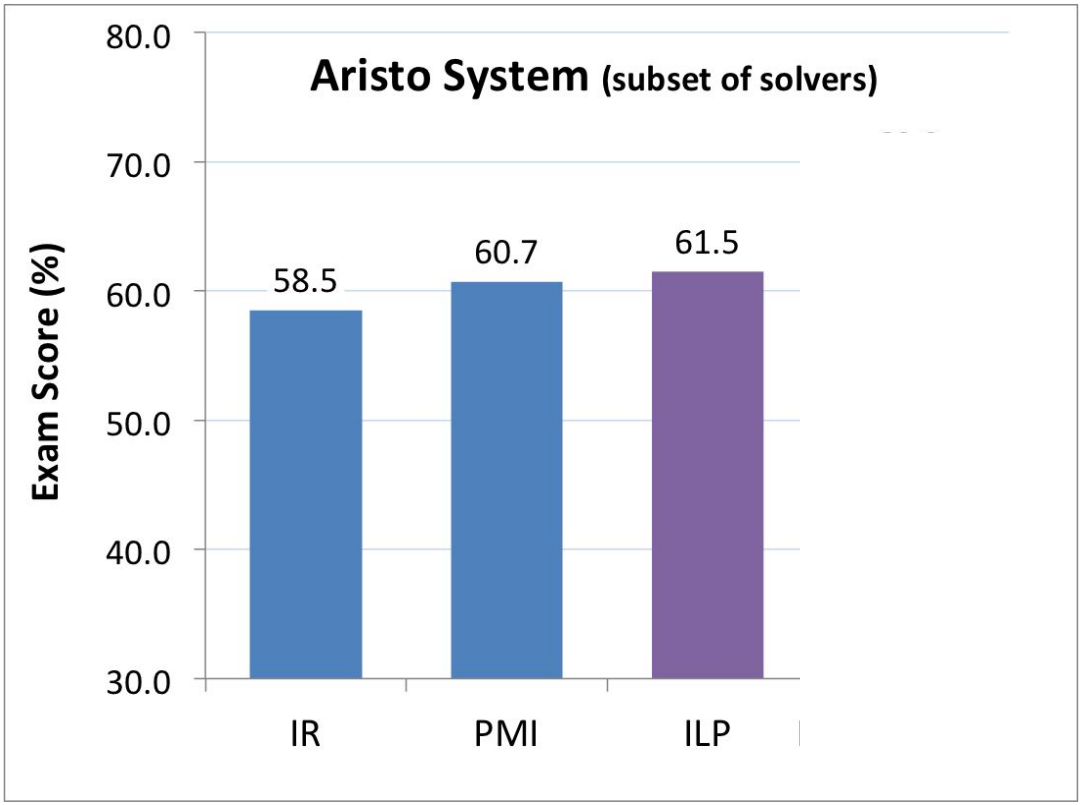
- **Baselines:**
  - IR Solver: Information Retrieval using Lucene search
    - Using 280 GB of plain text (50B tokens) "waterloo" corpus [AAAI, 2015]
    - IR Solver(tables): Using same tables as TableILP
  - PMI Solver: Statistical correlation using pointwise mutual info.
    - Using 280 GB of plain 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



Results



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, …)

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

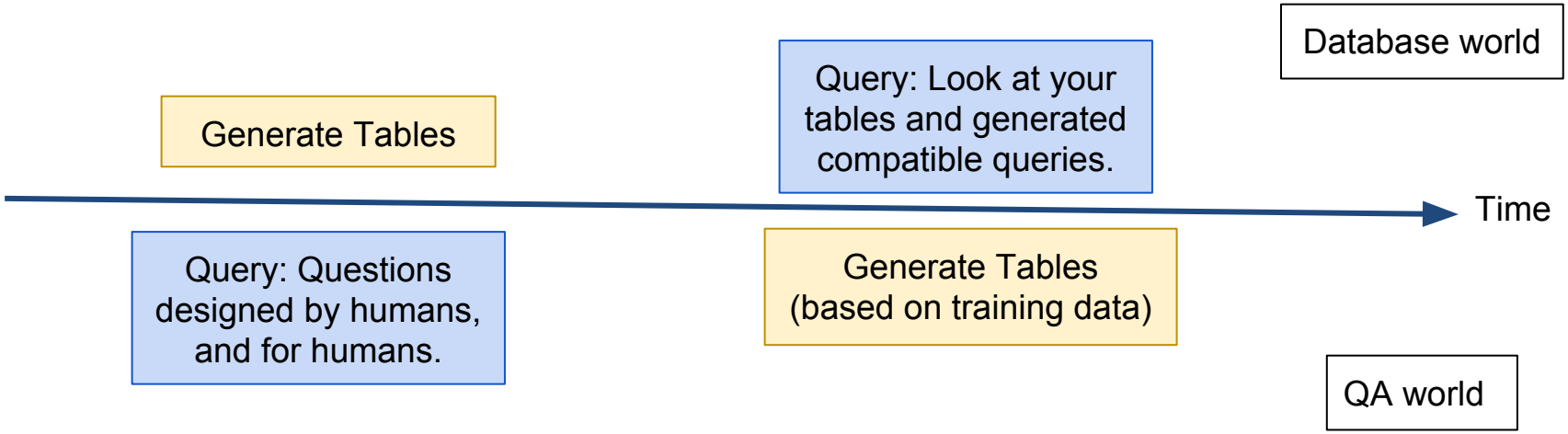
> In New York State, the longest period of daylight occurs during which month?
> (A) *eastern* (B) June (C) *history* (D) *years*

More experiments in the paper!

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

How does it compare to SQL query?

- Database technology has more than 40 years history
  - How are you different from that?

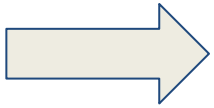


Beyond tables: Tuple Inference

TableILP

Small number of semi-structured rows

Khashabi et al IJCAI'16



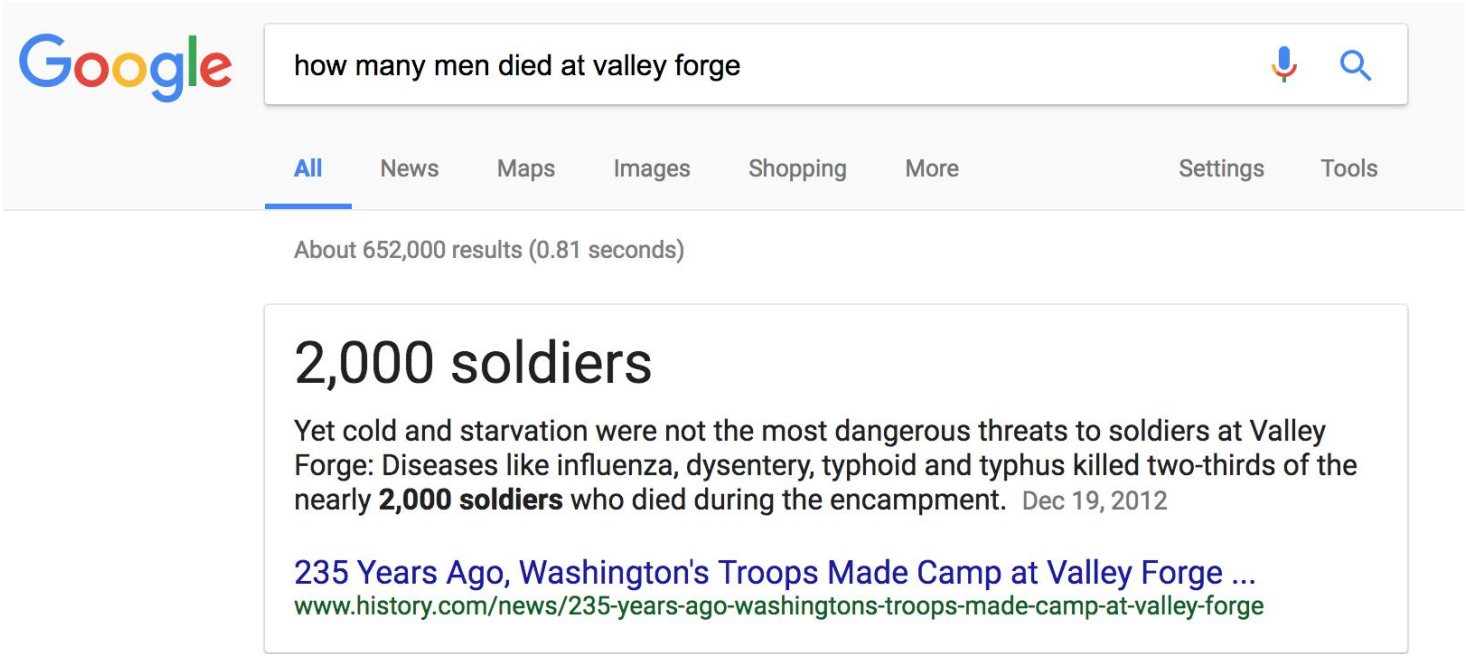
Tuple Inference

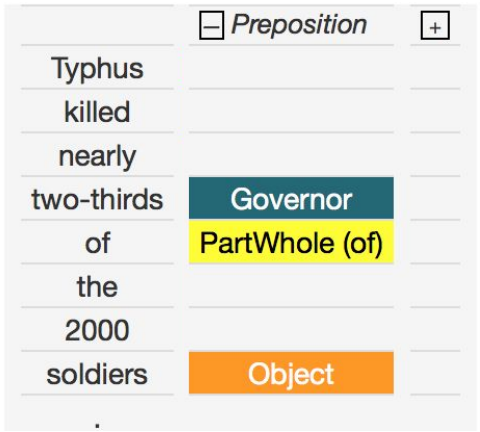
Large number of simple rows

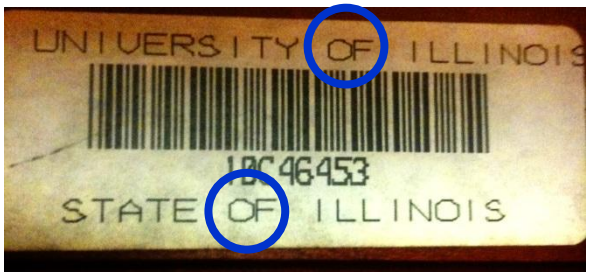
Khot et al, ACL'17

- Inference over **independent rows**
- **Auto-generated short triples** will often lose critical context
- Additional structure present in the subject/object phrases
- Scaling to millions of tuples
- Matching on 100s of relations (e.g., ~150 in animal tensor)

Beyond tables: modelling compositionality







Not all "of"s are the same

- One argument is a part of another.
  - The governor is a number
  - The object is a group modified by the governor.

Penn
University of Pennsylvania

Example

P: Teams are under pressure after PSG purchased Neymar this season. Chelsea purchased Morata. The Spaniard looked like he was set for a move to Old Trafford for the majority of the summer only for Manchester United to sign Romelu Lukaku instead, paving the way for Morata to finally move to Chelsea for an initial £56m.

Q: *Who did Chelsea purchase this season?*

A: *{Alvaro Morata, Neymar, Romelu Lukaku}*

Example

P: Teams are under pressure after PSG purchased Neymar this season. Chelsea purchased Morata. The Spaniard looked like he was set for a move to Old Trafford for the majority of the summer only for Manchester United to sign Romelu Lukaku instead, paving the way for Morata to finally move to Chelsea for an initial £56m.

Q: *Who did Chelsea purchase this season?*

A: *{Alvaro Morata, Neymar, Romelu Lukaku}*

Simple "lookup" based on proximity to question words, answer type

- Basic word overlap suffices
- Neural methods (e.g., BiDAF) excel at

Example, Rephrasing

P: Teams are under pressure after PSG purchased Neymar this season. Morata, the recent acquisition by Chelsea, will start for the team tomorrow. The Spaniard looked like he was set for a move to Old Trafford for the majority of the summer only for Manchester United to sign Romelu Lukaku instead, paving the way for Morata to finally move to Chelsea for an initial £56m.

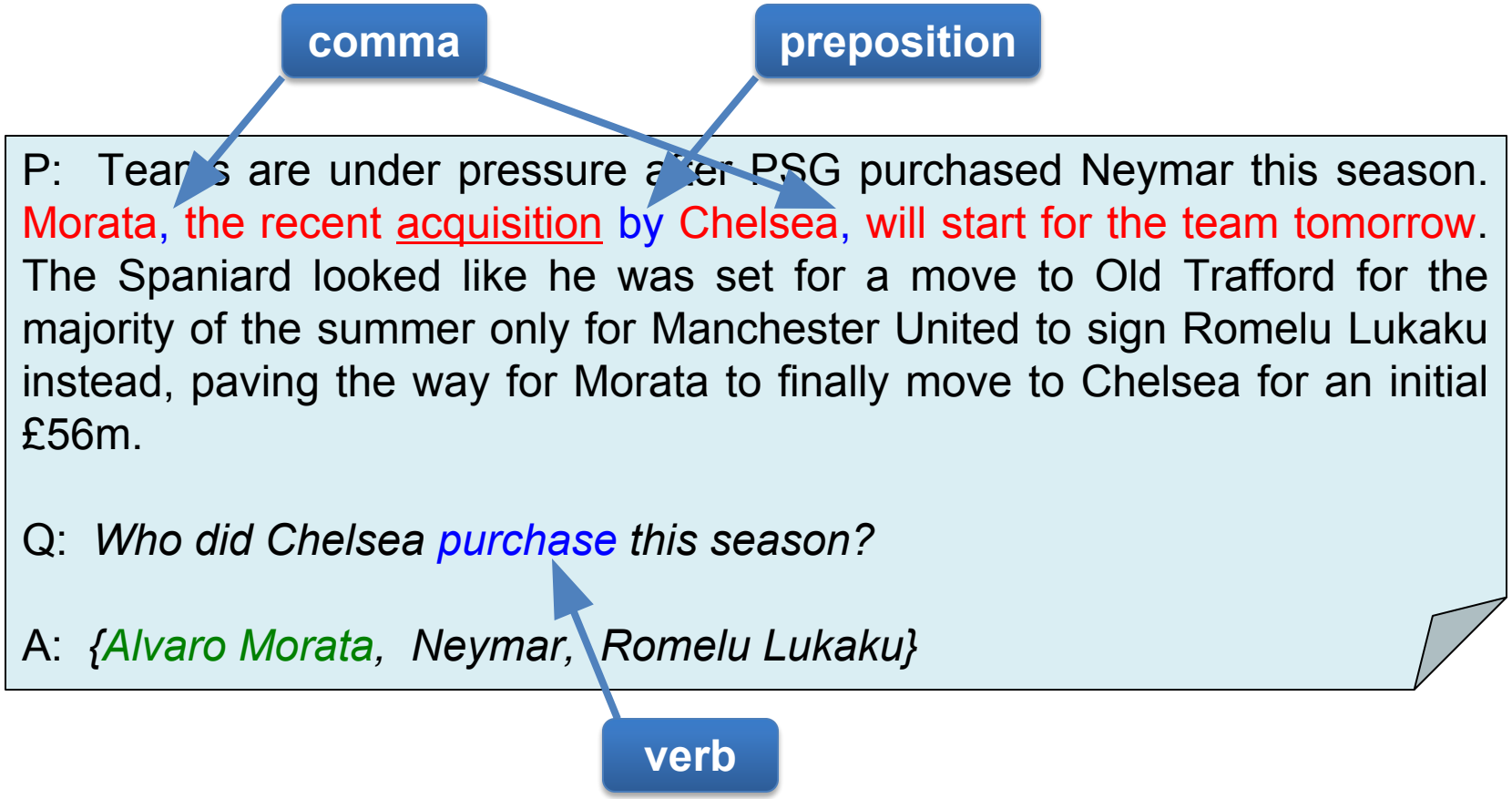
Q: *Who did Chelsea purchase this season?*

A: *{Alvaro Morata, Neymar, Romelu Lukaku}*

Simple rewording can confuse solvers

- E.g., BiDAF outputs *"Neymar"*

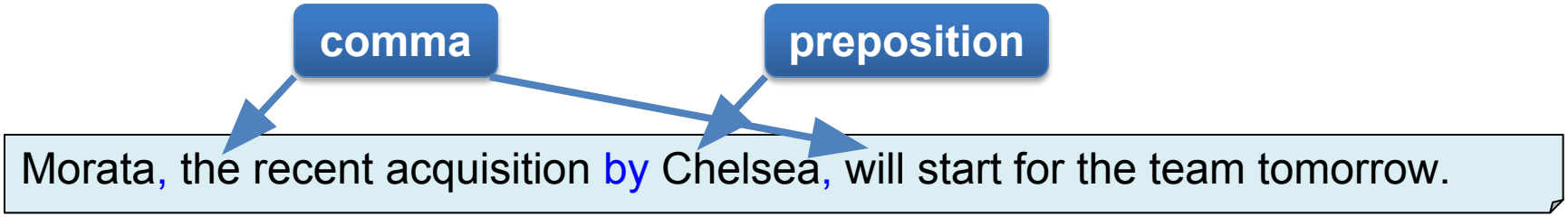
Example, Rephrasing



Linguistic understanding can help!

- Verbs, preposition, punctuation
- Domain agnostic => can use pre-trained NLP modules

Example, Rephrasing, simplified



- **Prepositional predicate: by (agent)**
  - The action done: the recent acquisition
  - Who/what did the action: Chelsea
- **Comma predicate: , (substitute)**
  - indicates an apposition structure

Identify the relation expressed by the predicate, and its arguments

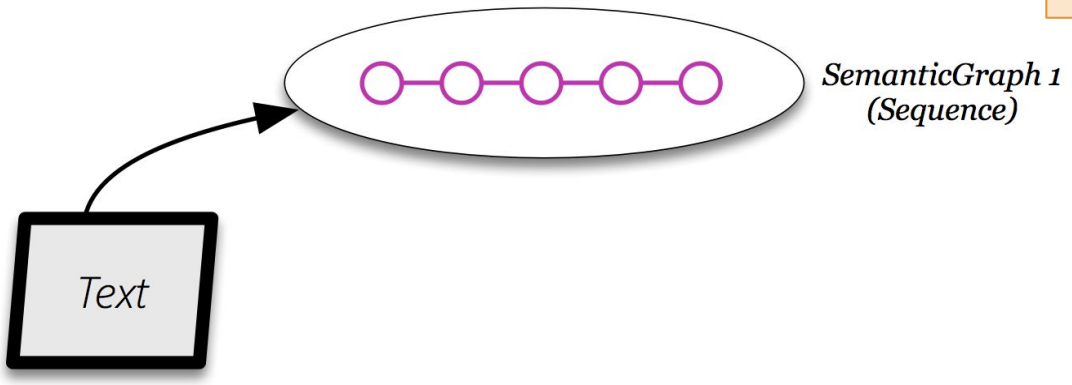
- **Verb predicate: purchase (agent)**
  - Purchaser (A0): Chelsea
  - Thing purchased (A1): Who

Linguistic understanding can help!

Create a unified representation as a family of graphs

- predicate-argument, trees, clusters, sequences

A single representation is not enough to capture complexity of language

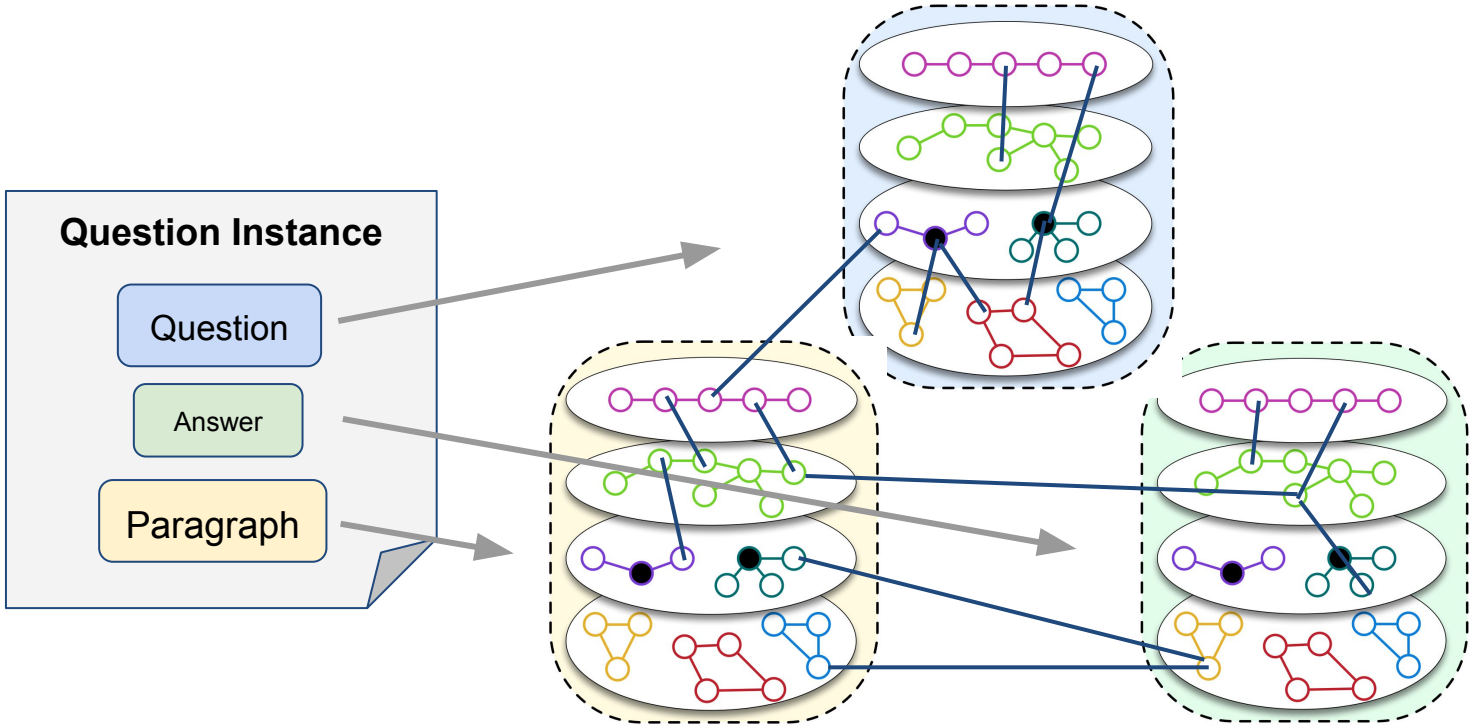


Instantiate with off the shelf NLP annotators

- Verb-SRL, Comma-SRL, Nom-SRL, Prep-SRL, Coref

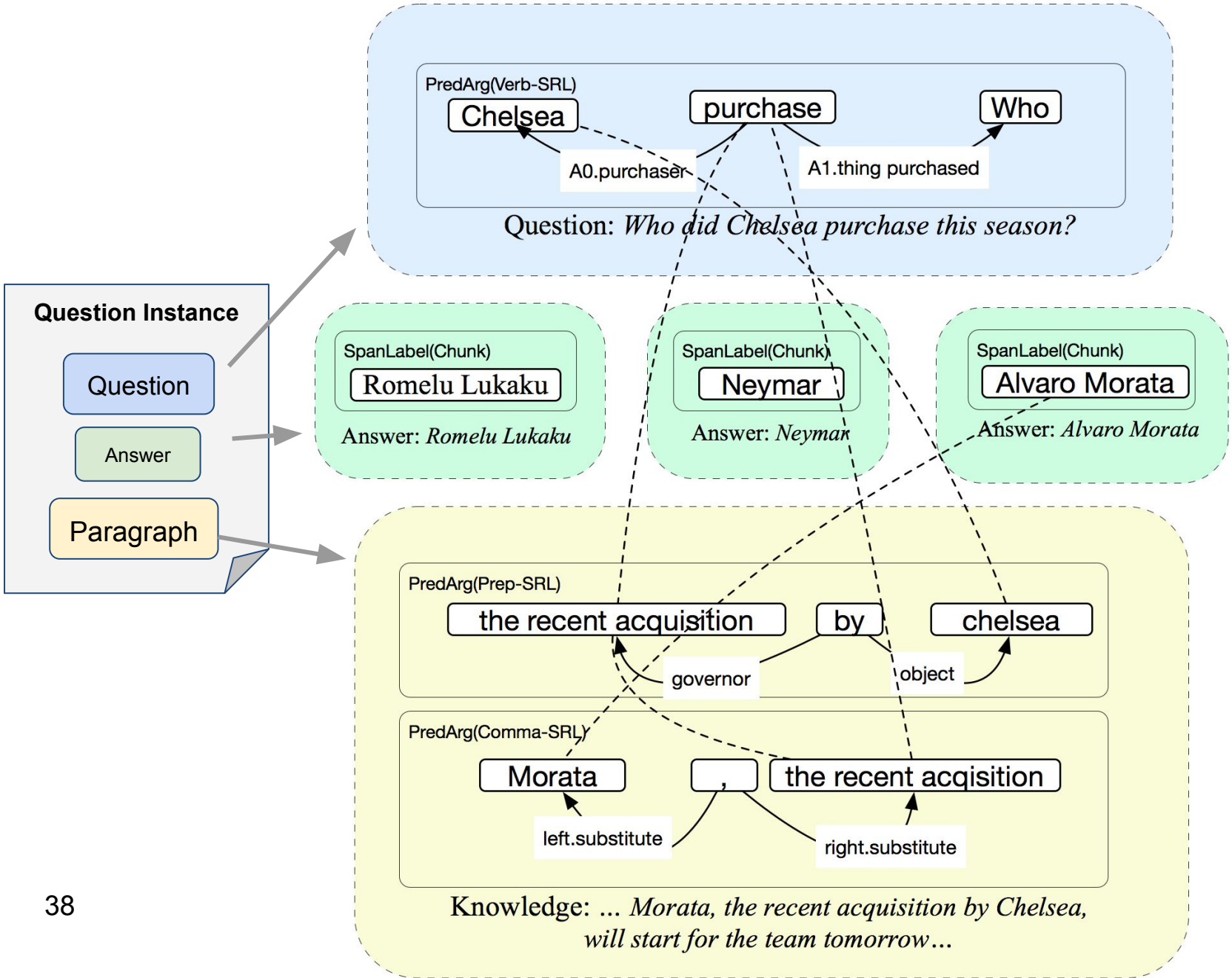
Bridges between graphs

- **Augmented Graph** is the graph which contains potential alignments between elements of any two graphs



- Connections via similarity / entailment
- Surface form well as semantics of their labels

SemanticILP: Example



SemanticILP

Translate QA into a search for an optimal subgraph

Constraint: Incorporate **global** and **local** constraints

- **Global** e.g.
  - Have ends in question, paragraph and an answer
  - Connected graph
  - Exactly one answer
- **Local** e.g.
  - If using a pred-arg graphs, use at least predicate and argument
  - If using a coref connected-comp. use at least two nodes

Objective: Capture what's a valid reasoning, what's preferred

- **Preferences** e.g.
  - Use less sentences
  - Use sentences nearby
  - If using a pred-arg graph, give priority to the subject

Formulate as Integer Linear Program (ILP) optimization

- Solution points to the best supported answer

Results #1: Aristo Questions

- Input: **Science question** Q with 4 answer options A
- Text: paragraph P obtained by concatenating top k Lucene-retrieved sentences for various answer options

Dataset
Regents 4th
Public 4th
Regents 8th
Public 8th

(exam scores, shown as a percentage)

Results #2: ProcessBank [EMNLP-2014]

- Input: **Biology question** with 2 answer options

Dataset	**BiDAF**	**BiDAF tuned**	**IR**	**SyntProx Baseline***	**ProRead*** (structural supervision)	**SemanticILP** (linear comb. of components)
Process Bank**	58.7	61.3	63.8	61.9	68.1	67.9

SemanticILP does not rely on domain-specific process structure annotation

- Close to the specialized, state-of-the-art ProRead system
- Substantially better than syntax-based and neural baselines

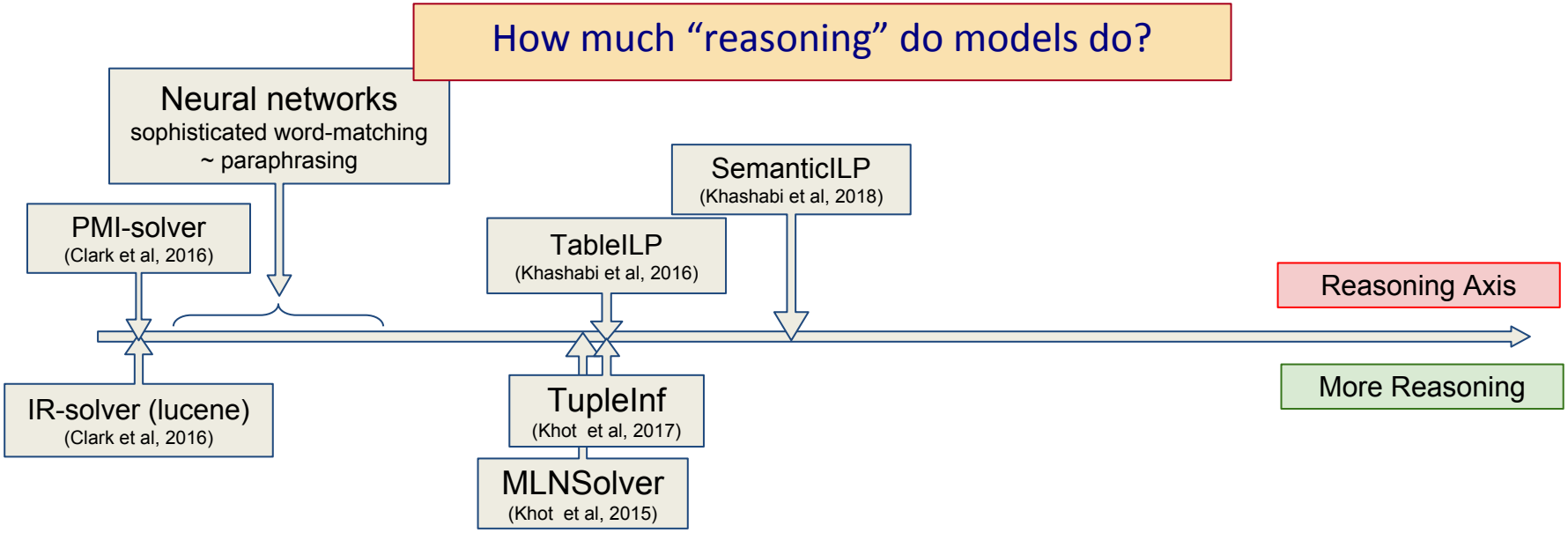
One single system tested on different datasets.

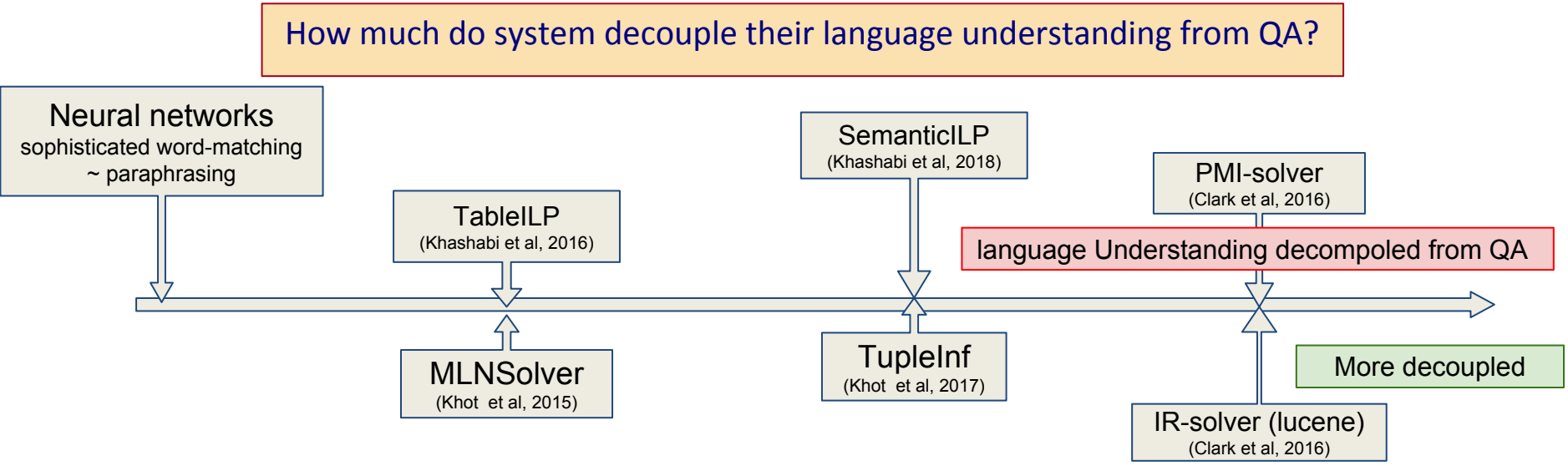
More experiments in the paper!

* Berant et al. (EMNLP, 2014)

** ~70% of the original dataset; true/false and temporal questions currently out of scope

Where are we?





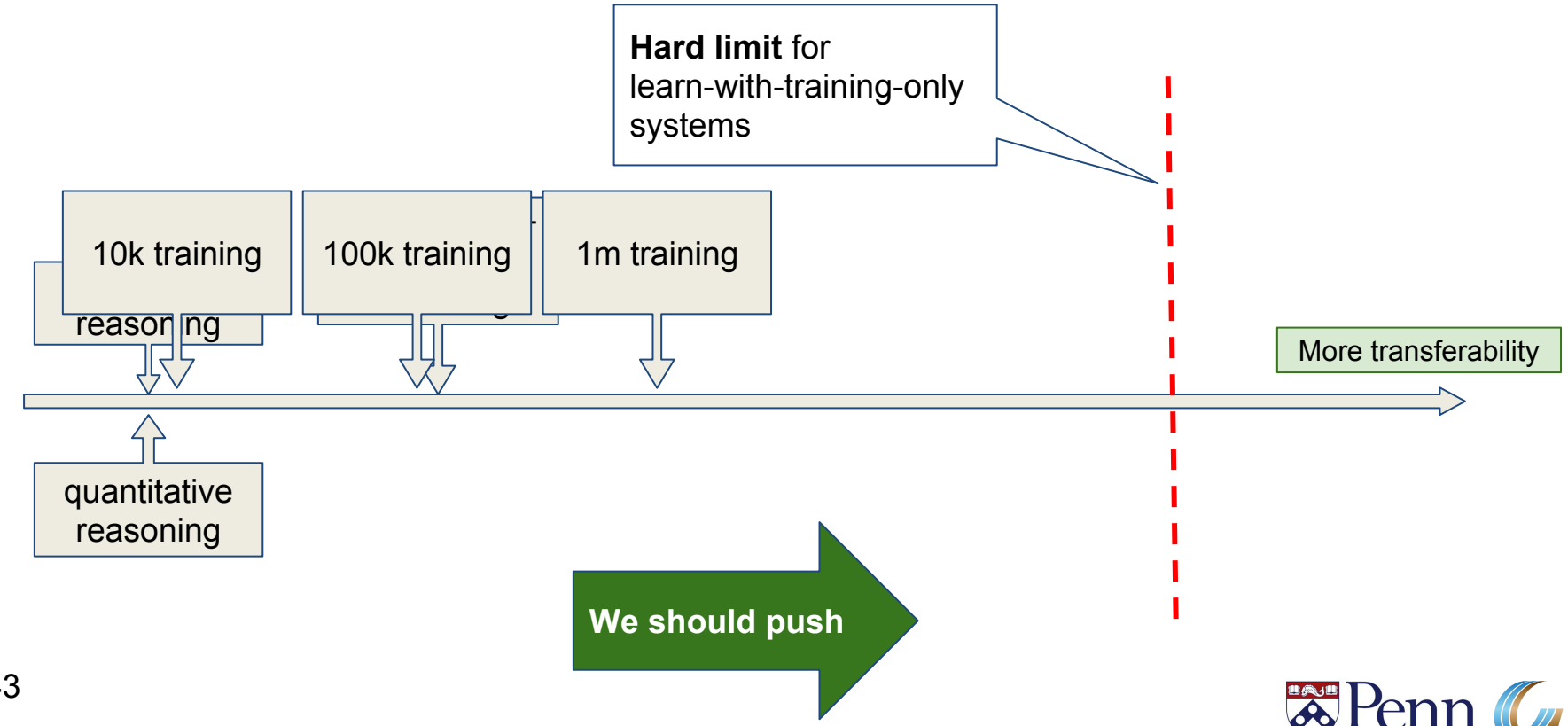
Transferability

For a “good” QA there is no notion of domain or dataset.

- Receive a question and give an answer

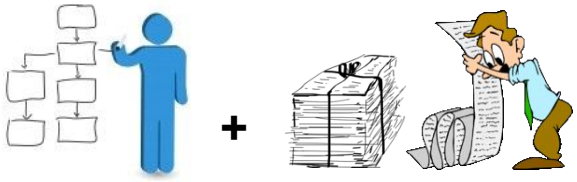
Many factors

- Reasoning shouldn’t be defined too narrowly
- Language understanding should be equated with training on datasets.



Takeaways

1. **Semi-structured inference**
   - We showed that can be very effective & robust
   - Goes beyond factoid-style QA
     - Chaining facts
     - First QA system to combine multiple semantic abstractions
     - State-of-the-art results on multiple datasets with different characteristics



2. **Reasoning** is a key to progress in QA.
   - No universal magic box
   - Decoupling QA and Language Understanding
   - Transferability is a key challenge

Advertisement (I)

CogCompNLP most extensive NLP annotator

Many annotators: POS, NER, verb-SRL, comma-SRL, etc.
Code: https://github.com/CogComp/cogcomp-nlp

		Sentence splitting	Tokenizing	Lemmatizing	Part of Speech tagging	Chunking	NER (4 labels)	Extended NER (18 labels)	Constituency Parsing	Dependency Parsing	Quantity Normalization	Verb-sense Classification	Temporal Normalization	Mention Detection	Comma SRL	Preposition SRL	Propbank (verb) SRL	Nombank (nominal) SRL	Coreference resolution	Relation Extraction	Sentiment	OpenIE
	CogCompNLP (ours)	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Java	CoreNLP	✓	✓	✓	✓	✓	✓		✓	✓				✓					✓	✓	✓	✓
Java	OpenNLP	✓	✓	✓	✓	✓	✓		✓										✓			
Python	SpaCy	✓	✓	✓	✓	✓	✓			✓												
Python	NLTK	✓	✓	✓	✓	✓	✓															
Python	TextBlob	✓	✓	✓	✓	✓															✓	

Advertisement (II)

CogCompNLPy Light-weight Python NLP annotators

Many annotators: POS, NER, verb-SRL, comma-SRL, etc.
Code: https://github.com/CogComp/cogcomp-nlpy

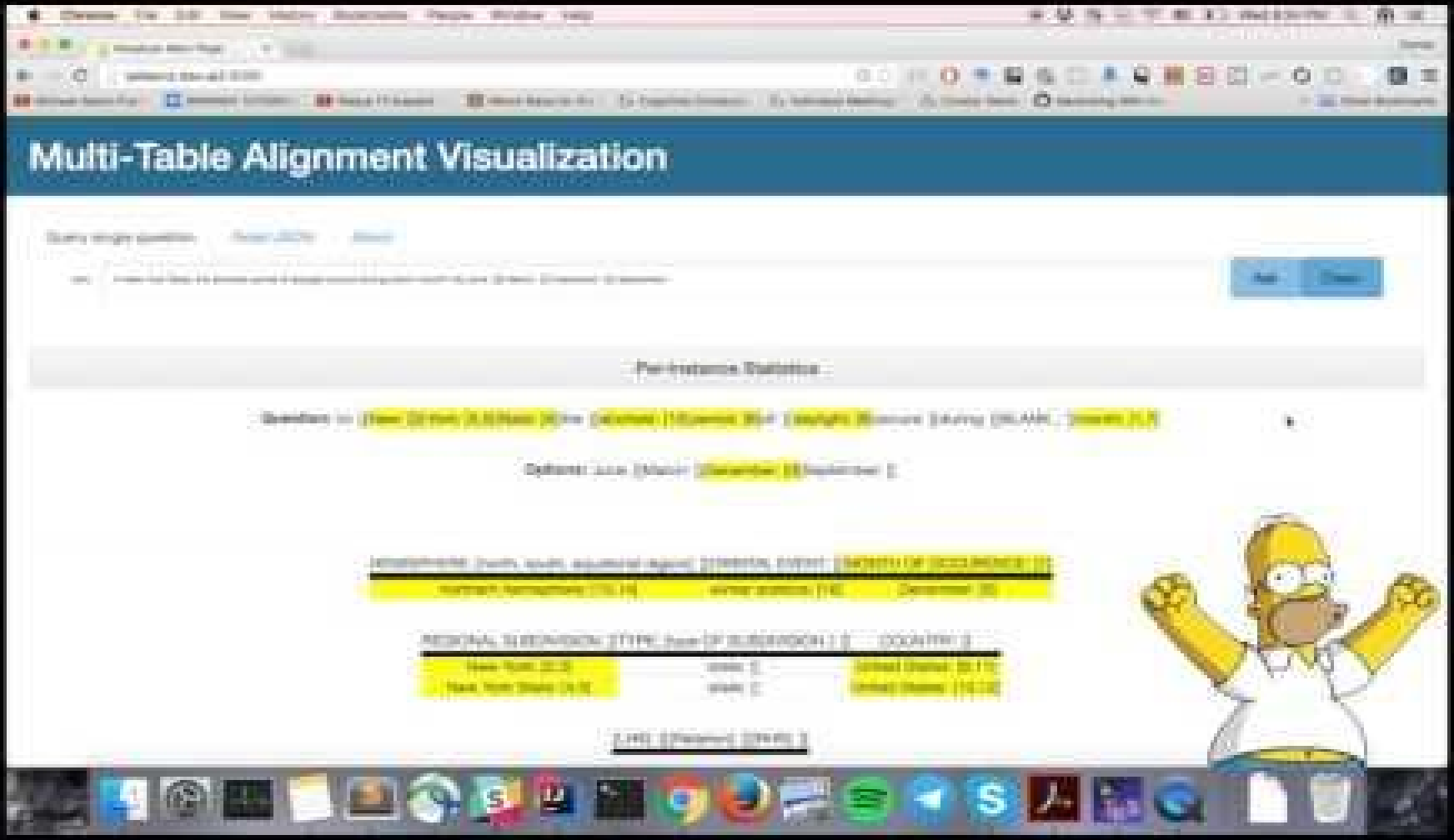
```
from sioux import remote_pipeline
pipeline = remote_pipeline.RemotePipeline()
doc = pipeline.doc("Hello, how are you. I am doing fine")

print(doc.get_lemma)
# will produce (hello Hello) (, ,) (how how) (be are) (you you) (. .) (i I) (be
am) (do doing) (fine fine)

print(doc.get_pos)
# will produce (UH Hello) (, ,) (WRB how) (VBP are) (PRP you) (. .) (PRP I) (VBP
am) (VBG doing) (JJ fine)
```

EXTRA SLIDES

Demo



Motivation: Three Challenges

- Diverse linguistic constructs make QA systems brittle
  - ⇒ Even the best systems are easily fooled by simple textual variations
- Limited training data in "interesting" QA domains
  - ⇒ Paradigm of learning everything end-to-end doesn't seem viable
- Limited question understanding in Aristo solvers
  - ⇒ knowledge: explored several representations
  - ⇒ question: still treated as tokens/chunks

Goal: Address these in the context of multiple-choice questions with supporting text, by reasoning over semantic abstractions of text

Question Answering Now

- Majority of QA systems do “sophisticated” paraphrasing.
  - Not much reasoning
- An experiment on a recent popular dataset SQuAD (Rajpurkar et al, 2016):

Experiment

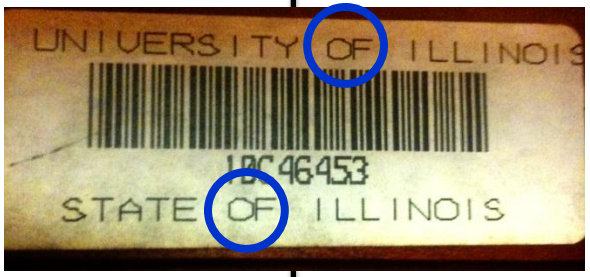
- Take 50 questions and the paragraph contained its answer.
- Break the paragraph into sentences.
- Create (question, sentence) pairs, for any sentence in the paragraph
- For each (question, sentence) ask 3 people whether they can answer the question.
- Repeated it for all the sentences in each paragraph.

Result

- At least 2 (out of 3) people said they can answer 74% of questions, given a single sentence
- Manual inspection: The remaining questions all required co-reference reasoning.

Computational Questions

- *John, a fast-rising politician, slept on the train to Chicago.*
- **Verb Predicate: sleep**
  - Sleeper: John, a fast-rising politician
  - Location: on the train to Chicago
- **Who was John?**
  - Relation: Apposition (comma)
  - John, a fast-rising politician
- **What was John's destination?**
  - Relation: Destination (preposition)
  - train to Chicago

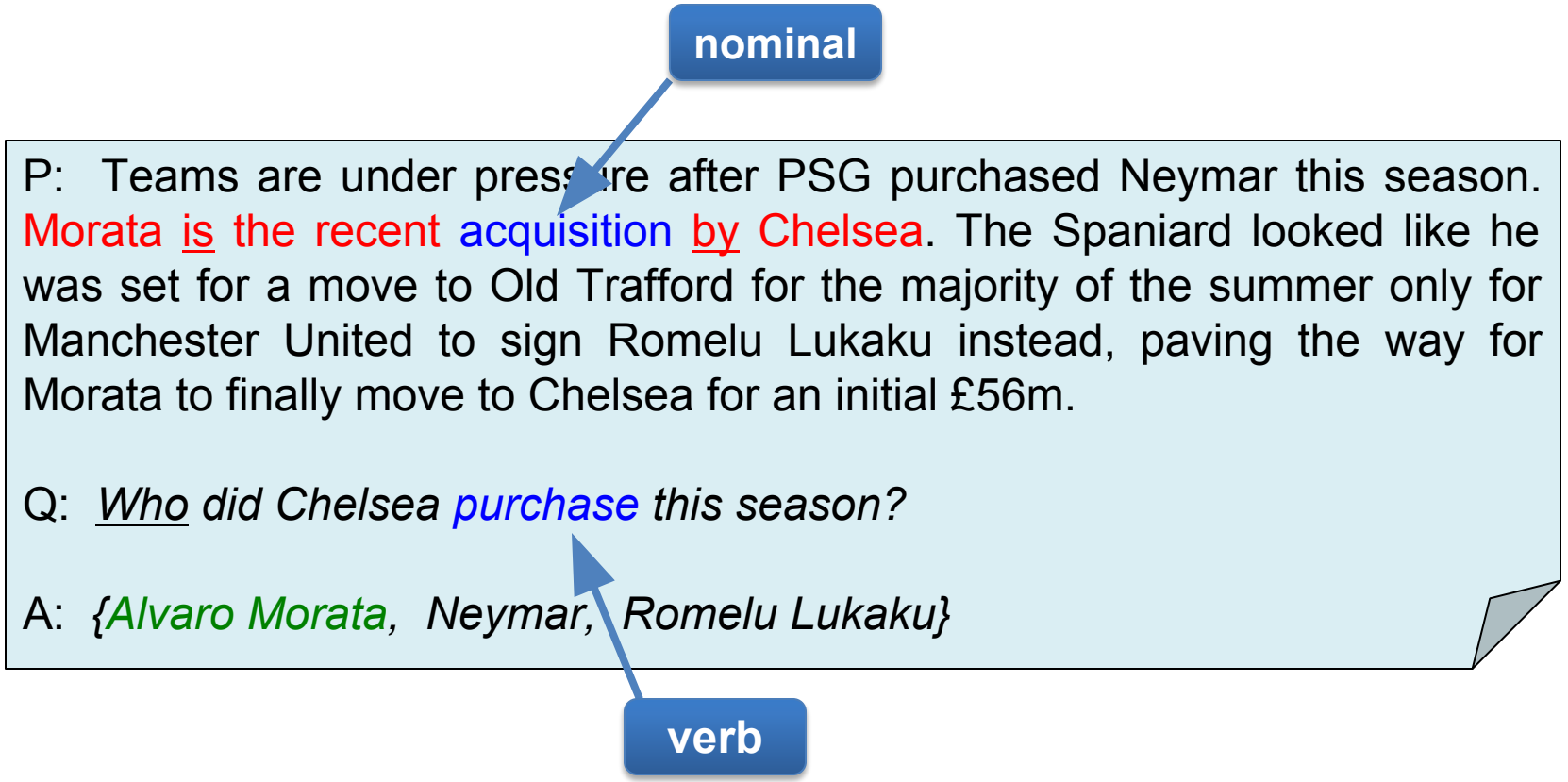


Identify the relation expressed by the predicate, and its arguments

Extended Semantic Role Labeling

- Improved sentence level analysis; dealing with more phenomena.
  - Semantic role labelling
  - Events, Entailment, Winograd schemas
- Abstract Meaning Representation (AMR) [Banarescu et al. 2013]
  - Expensive to produce large amounts of hand-annotated AMRs
  - Especially for other languages/genres†
  - Limitations in terms of phenomena covered (hard to add more)

Example, Rephrased



Linguistic understanding can help!

- Verbs and their nominalization
- Domain agnostic => can use pre-trained NLP modules

Example, Rephrased

P: Teams are under pressure after PSG purchased Neymar this season. Morata is the recent acquisition by Chelsea. The Spaniard looked like he was set for a move to Old Trafford for the majority of the summer only for Manchester United to sign Romelu Lukaku instead, paving the way for Morata to finally move to Chelsea for an initial £56m.

Q: *Who did Chelsea purchase this season?*

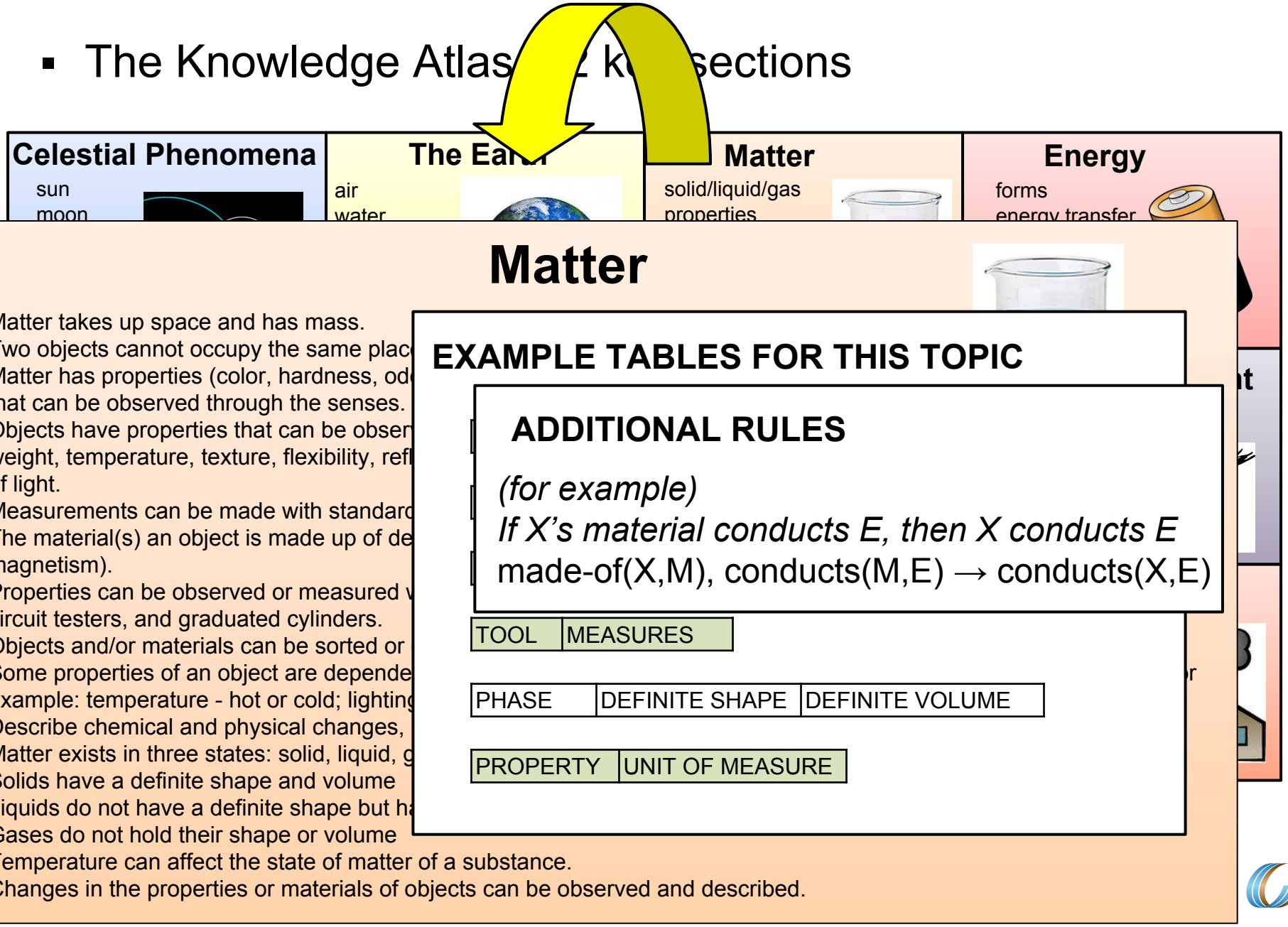
A: *{Alvaro Morata, Neymar, Romelu Lukaku}*

Simple rewording can confuse solvers

- E.g., BiDAF outputs *"Neymar this season. Morata"*

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 standard
The material(s) an object is made up of de
magnetism).
Properties can be observed or measured v
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; lighting
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 ha
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

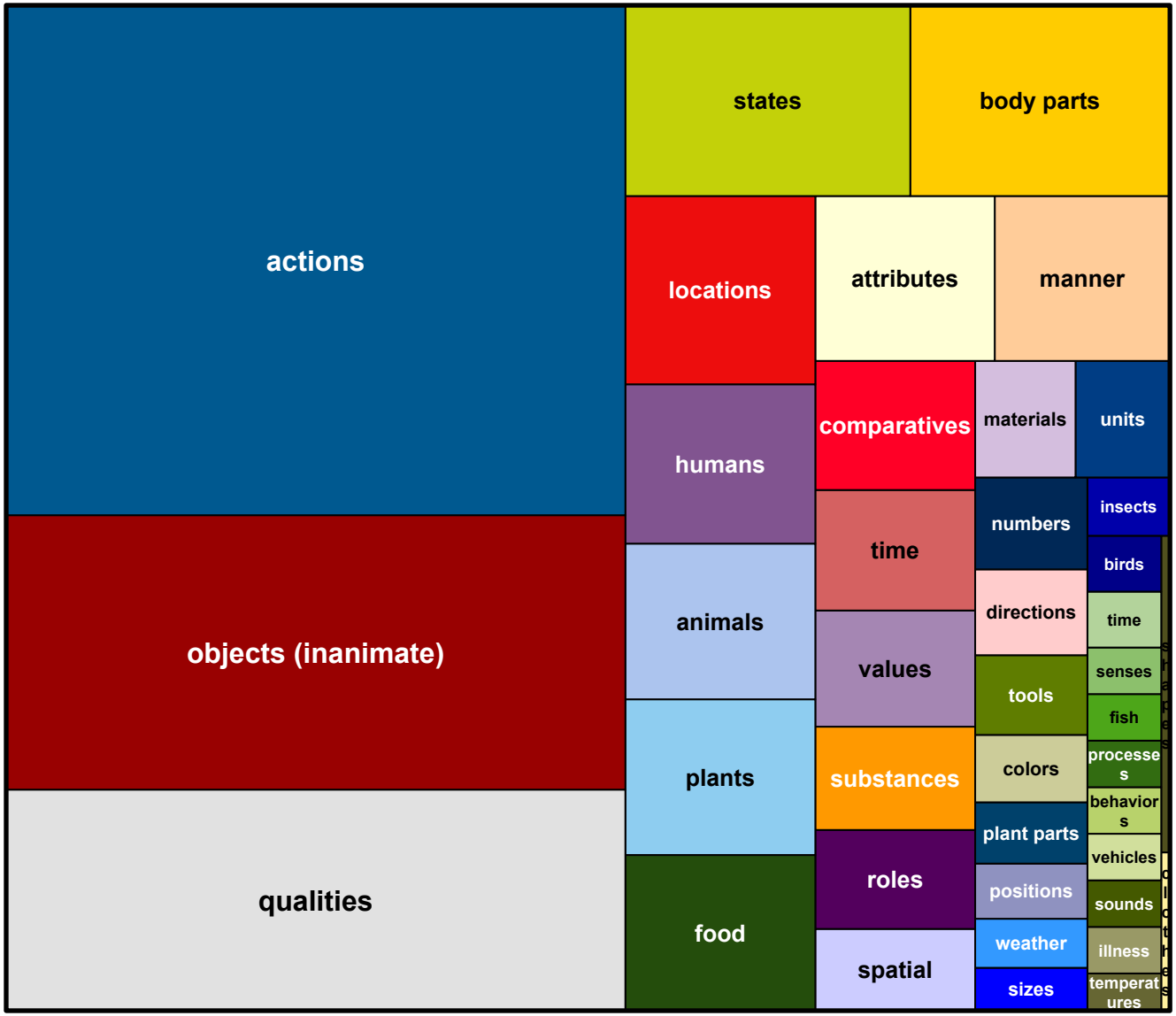
(for example)
If X's material conducts E, then X conducts E
made-of(X,M), conducts(M,E) → conducts(X,E)

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

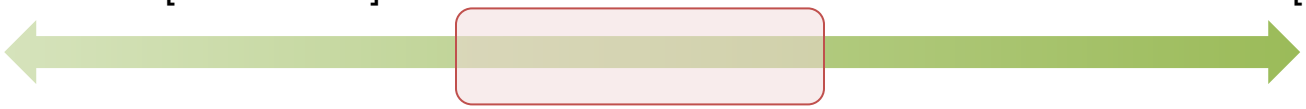
Penn
University of Pennsylvania

Semi-Structured Inference: Challenge #2

Reasoning: effective, controllable, scalable

RULE solver [AKBC 2014]

MLN solver [EMNLP 2015]



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

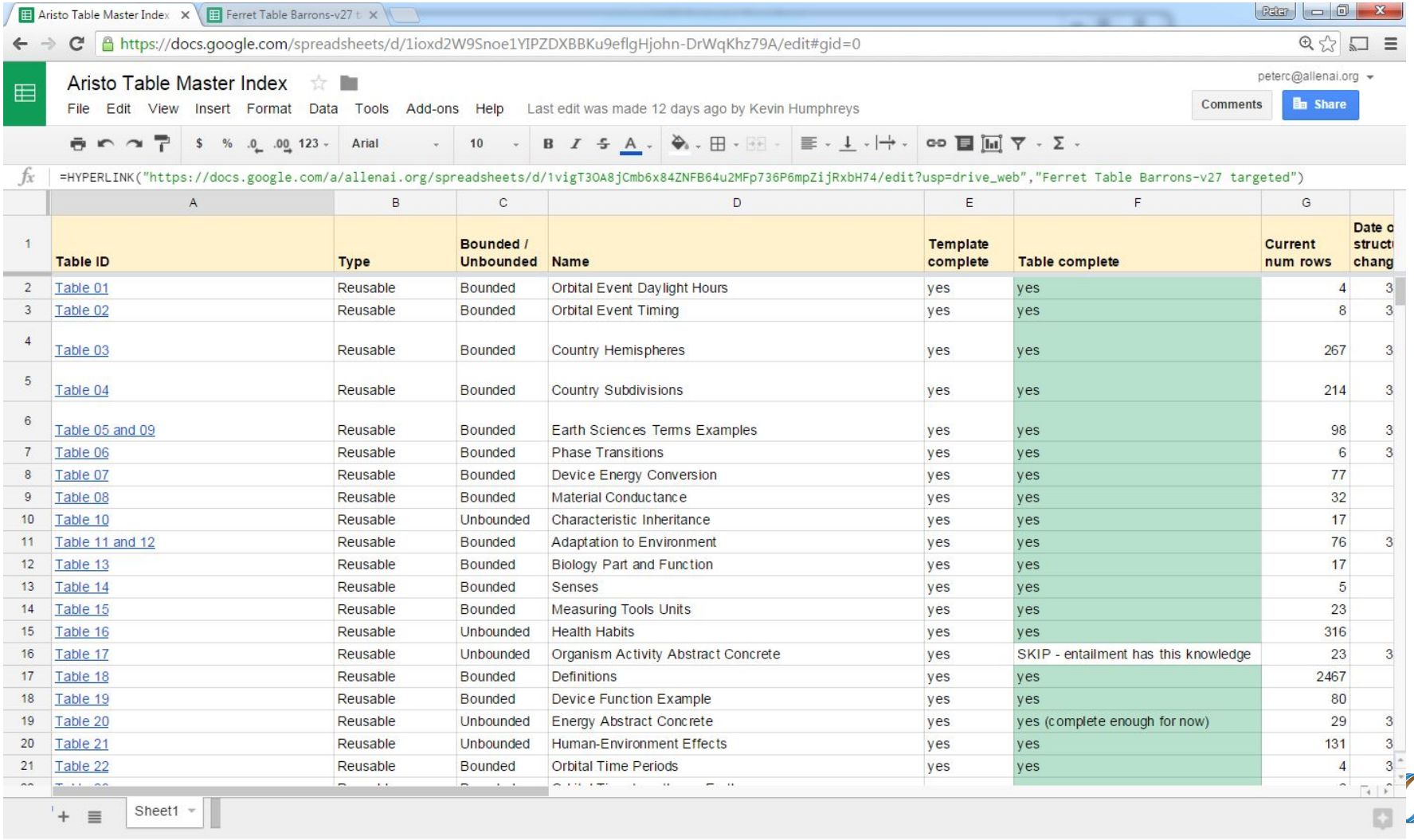
Evaluation: Ablation Study

- 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'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

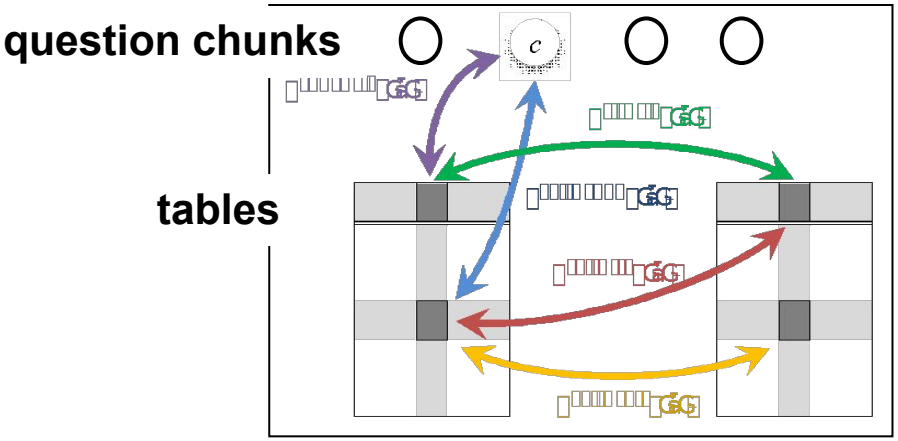
Category	Quantity	Average
ILP complexity	#variables	1043.8
	#constraints	4417.8
	#LP iterations	1348.9
Knowledge use	#rows	2.3
	#tables	1.3
Timing stats	model creation	1.9 sec
	solving the ILP	2.1 sec

- 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

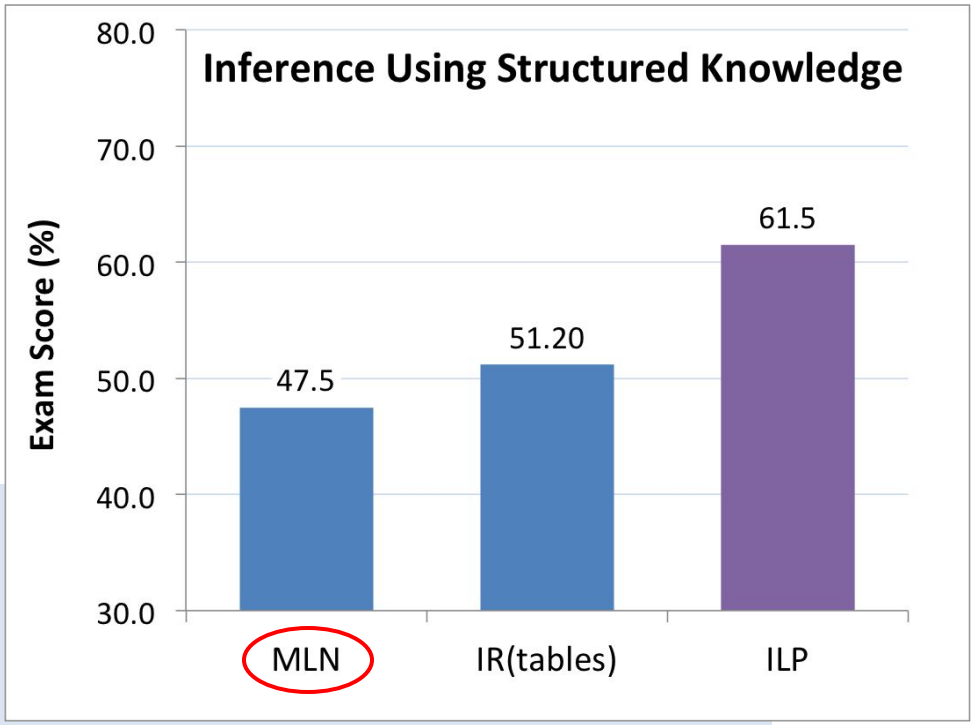
ILP Model: Constraints

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

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)

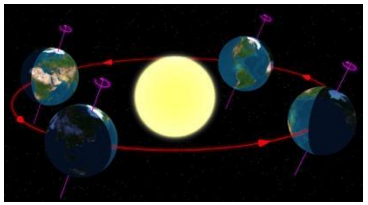
Two Approaches to Question Answering

New Zealand shortes 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

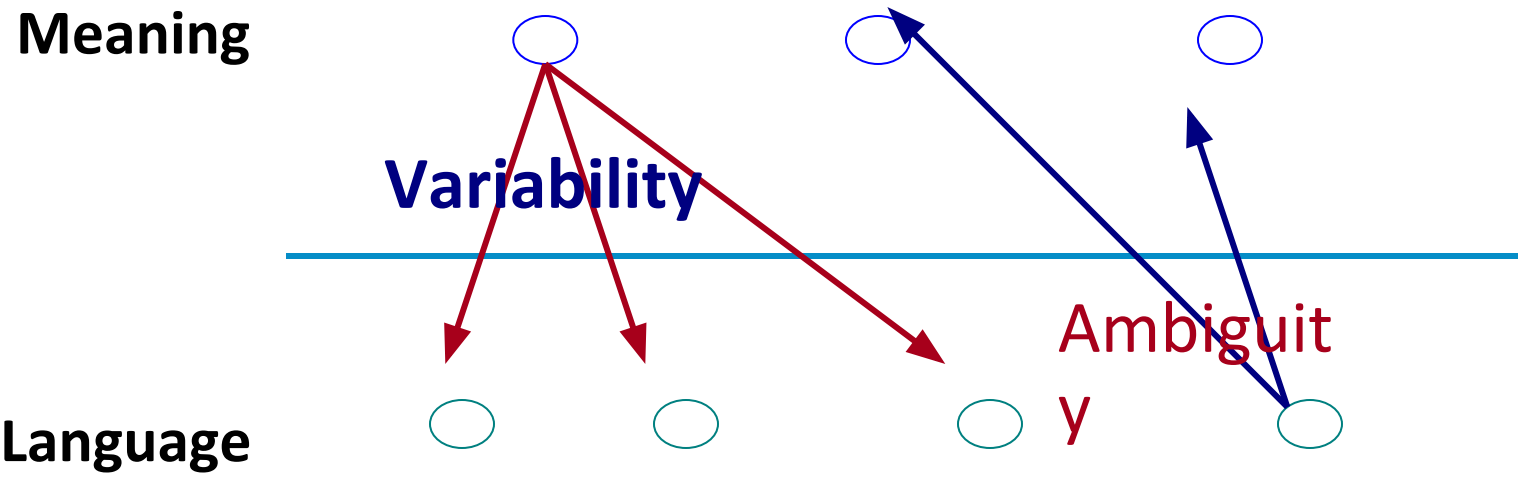


- **Information retrieval** / statistical association
  - ✓ easy, generalizes well, often effective
  - × limited to simple reasoning
  - × expects answers explicitly written somewhere

Why is it Difficult?

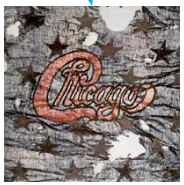
One cannot simply map natural language to a representation that gives rise to reasoning

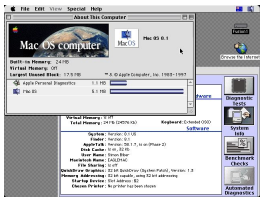
Midas: I hope that everything I touch becomes gold.



AMBIGUITY

It's a version of ***Chicago*** – the standard classic ***Macintosh*** menu font, with that distinctive thick diagonal in the "N".	***Chicago*** was used by default for ***Mac*** menus through ***MacOS 7.6***, and ***OS 8*** was released mid-1997..	***Chicago VIII*** was one of the early 70s-era ***Chicago*** albums to catch my ear, along with ***Chicago***





Variability in Natural Language Expressions

Determine if Jim Carpenter works for the government

Jim Carpenter works for the U.S. Government.

The American government employed Jim Carpenter.

Jim Carpenter was fired by the US Government.

Jim Carpenter worked in a number of important positions. …. As a press liaison for the IRS, he made contacts in the white house.

Russian interior minister Yevgeny Topolov met yesterday with his US counterpart, Jim Carpenter.

Former US Secretary of Defense Jim Carpenter spoke today…

Conventional programming techniques cannot deal with the variability of expressing meaning nor with the ambiguity of interpretation

Machine Learning is needed to support abstraction over the raw text, and deal with:

- Identifying/Understanding Relations, Entities and Semantic Classes
- Acquiring knowledge from external resources; representing knowledge
- Identifying, disambiguating & tracking entities, events, etc.
- Time, quantities, processes…

Where are we?

Aristo: Ensemble Approach

[AAAI-2016]