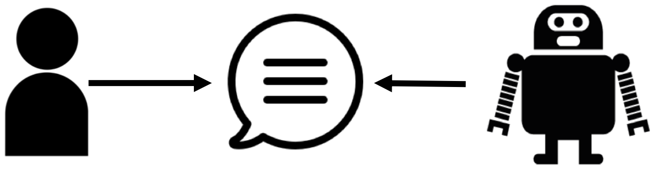
Reasoning-Driven Question Answering for Natural Language Understanding

Daniel Khashabi

Natural Language Understanding

Natural Language Understanding

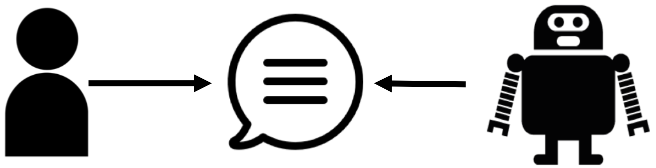
- Interpret a given text similar to humans.



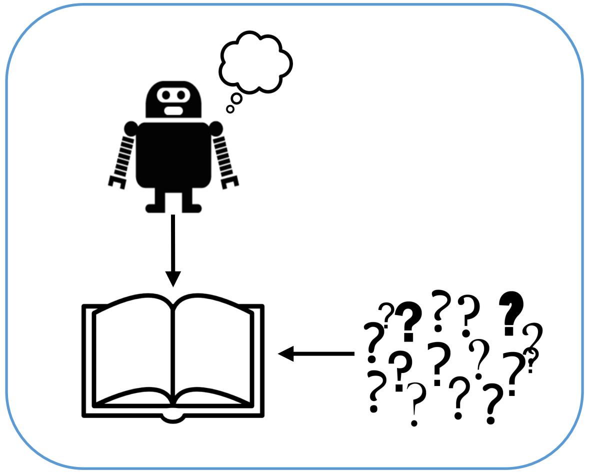
- Measuring the progress by answering questions.
  - A system that is better in understanding language, should have a higher chance of answering these questions.
  - This has been used in the field for many years [Winograd, 1972; Lehnert, 1977b; others]
    - Question Answering (QA),
    - Reading Comprehension (RC),
    - Textual Entailment (TE), etc.

Natural Language Understanding

- Interpret a given text similar to humans.

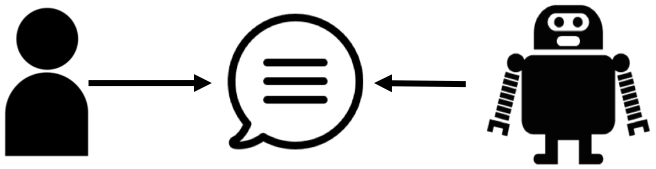


- Measuring the progress by answering questions.
  - A system that is better in understanding language, should have a higher chance of answering these questions.
  - This has been used in the field for many years [Winograd, 1972; Lehnert, 1977b; others]
    - Question Answering (QA),
    - Reading Comprehension (RC),
    - Textual Entailment (TE), etc.

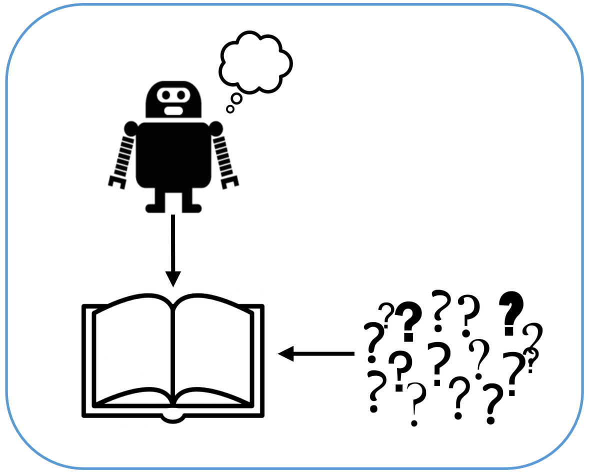


Natural Language Understanding

- Interpret a given text similar to humans.

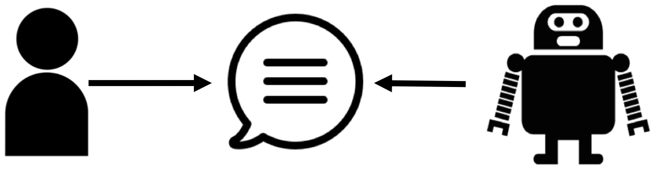


- Measuring the progress by answering questions.
  - A system that is better in understanding language, should have a higher chance of answering these questions.
  - This has been used in the field for many years [Winograd, 1972; Lehnert, 1977b; others]
    - Question Answering (QA),
    - Reading Comprehension (RC),
    - Textual Entailment (TE), etc.

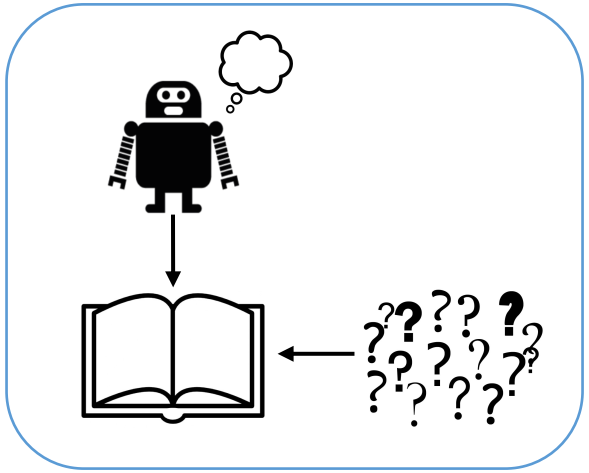


Natural Language Understanding

- Interpret a given text similar to humans.

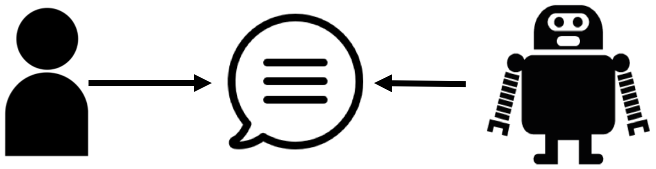


- Measuring the progress by answering questions.
  - A system that is better in understanding language, should have a higher chance of answering these questions.
  - This has been used in the field for many years [Winograd, 1972; Lehnert, 1977b; others]
    - Question Answering (QA),
    - Reading Comprehension (RC),
    - Textual Entailment (TE), etc.

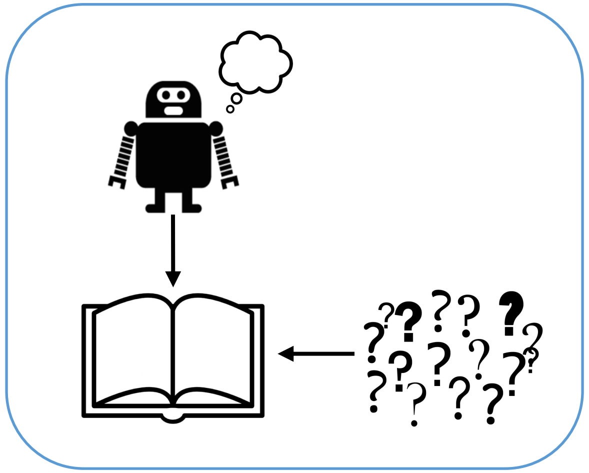


Natural Language Understanding

- Interpret a given text similar to humans.



- Measuring the progress by answering questions.
  - A system that is better in understanding language, should have a higher chance of answering these questions.
  - This has been used in the field for many years [Winograd, 1972; Lehnert, 1977b; others]
    - Question Answering (QA),
    - Reading Comprehension (RC),
    - Textual Entailment (TE), etc.



NLU Challenges: Ambiguity

- Making sense of strings.

The sentence from The New York Times

NLU Challenges: Ambiguity

- Making sense of strings.

"A 61-year old furniture salesman was pushed down the shaft of a freight elevator yesterday."

The sentence from The New York Times

NLU Challenges: Ambiguity

- Making sense of strings.

"A 61-year old furniture salesman was pushed down the shaft of a freight elevator yesterday."

The sentence from The New York Times

NLU Challenges: Ambiguity

- Making sense of strings.

"A 61-year old furniture salesman was pushed down the shaft of a freight elevator yesterday."

A 61-year old furniture salesman

The sentence from The New York Times

NLU Challenges: Ambiguity

- Making sense of strings.

"A 61-year old furniture salesman was pushed down the shaft of a freight elevator yesterday."

✓

A 61-year old furniture salesman

The sentence from *The New York Times*

NLU Challenges: Ambiguity

- Making sense of strings.

"A 61-year old furniture salesman was pushed down the shaft of a freight elevator yesterday."

A 61-year old furniture salesman

The sentence from The New York Times

NLU Challenges: Ambiguity

- Making sense of strings.

"A 61-year old furniture salesman was pushed down the shaft of a freight elevator yesterday."

✓

✗

A 61-year old furniture salesman

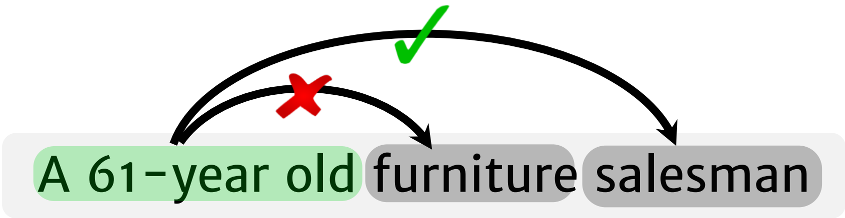
An antique furniture salesman

The sentence from The New York Times

NLU Challenges: Ambiguity

- Making sense of strings.

"A 61-year old furniture salesman was pushed down the shaft of a freight elevator yesterday."



An antique furniture salesman

The sentence from The New York Times

NLU Challenges: Ambiguity

- Making sense of strings.

"A 61-year old furniture salesman was pushed down the shaft of a freight elevator yesterday."

A 61-year old furniture salesman

An antique furniture salesman

The sentence from *The New York Times*

NLU Challenges: Ambiguity

- Making sense of strings.

"A 61-year old furniture salesman was pushed down the shaft of a freight elevator yesterday."

A 61-year old furniture salesman

An antique furniture salesman

The sentence from The New York Times

NLU Challenges: Ambiguity

- Making sense of strings.

"A 61-year old furniture salesman was pushed down the shaft of a freight elevator yesterday."

A 61-year old furniture salesman

An antique furniture salesman

The sentence from The New York Times

NLU Challenges: Ambiguity

- Making sense of strings.

"A 61-year old furniture salesman was pushed down the shaft of a freight elevator yesterday."

A 61-year old furniture salesman

An antique furniture salesman

shaft of a freight elevator

part-whole

The sentence from The New York Times

NLU Challenges: Ambiguity

- Making sense of strings.

"A 61-year old furniture salesman was pushed down the shaft of a freight elevator yesterday."

A 61-year old furniture salesman

An antique furniture salesman

~~shaft~~ of a freight elevator

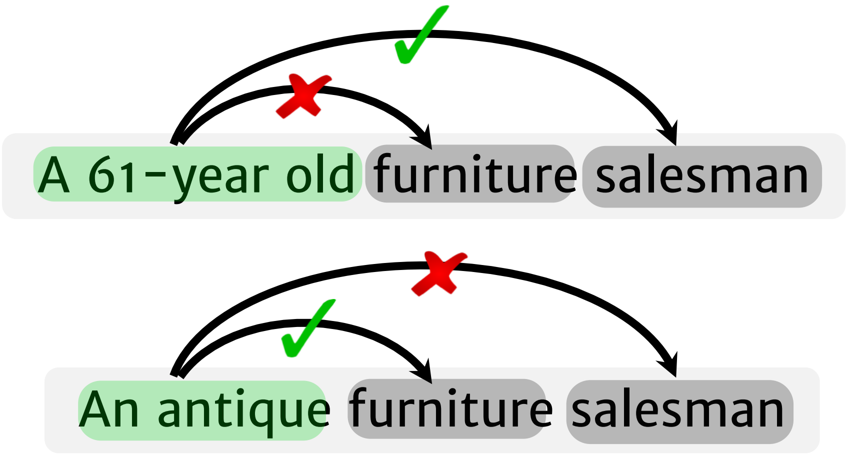
part-whole

The sentence from The New York Times

NLU Challenges: Ambiguity

- Making sense of strings.

"A 61-year old furniture salesman was pushed down the shaft of a freight elevator yesterday."



~~shaft~~ of a freight elevator	part-whole
height of a freight elevator	attribute

The sentence from The New York Times

NLU Challenges: Ambiguity

- Making sense of strings.

"A 61-year old furniture salesman was pushed down the shaft of a freight elevator yesterday."

A 61-year old furniture salesman

An antique furniture salesman

~~shaft~~ of a freight elevator ← part-whole

height of a freight elevator ← attribute

Oxford English Dictionary lists
<u>10</u> primary meanings for "of".

The sentence from The New York Times

NLU Challenges: Variability

The story from The New York Times

NLU Challenges: Variability

"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. Mr. Hug was pinned in the shaft for about half an hour until his cries attracted the attention of a porter."

The story from The New York Times

NLU Challenges: Variability

"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. Mr. Hug was pinned in the shaft for about half an hour until his cries attracted the attention of a porter."

The story from The New York Times

NLU Challenges: Variability

- A single meaning mentioned in many different ways.

"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. Mr. Hug was pinned in the shaft for about half an hour until his cries attracted the attention of a porter."

- Even more variability in bigger units (phrases, sentences, paragraphs, etc.)

The story from The New York Times

NLU Challenges: Reading Between the Lines

The sentence from The New York Times

NLU Challenges: Reading Between the Lines

- Lots of understanding is <u>only</u> implied from text.

    - *the car is significantly heavier than the man;*
    - *he had nowhere to go at the bottom of the pit;*
    - *if he didn't flatten himself, he would have died;*
    - *...*

The sentence from The New York Times

NLU Challenges: Reading Between the Lines

- Lots of understanding is <u>only</u> implied from text.

"The car stopped about 12 inches above him as he flattened himself at the bottom of the pit."

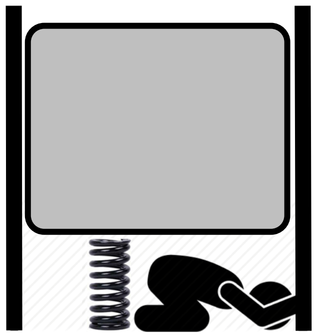
  - *the car is significantly heavier than the man;*
  - *he had nowhere to go at the bottom of the pit;*
  - *if he didn't flatten himself, he would have died;*
  - *...*

The sentence from The New York Times

NLU Challenges: Reading Between the Lines

- Lots of understanding is <u>only</u> implied from text.

> "The car stopped about 12 inches above him as he flattened himself at the bottom of the pit."



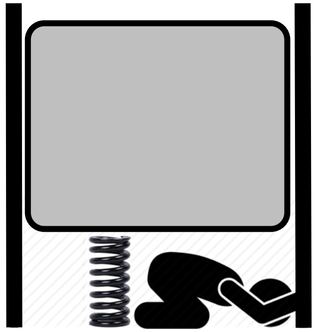
- *the car is significantly heavier than the man;*
- *he had nowhere to go at the bottom of the pit;*
- *if he didn't flatten himself, he would have died;*
- *...*

The sentence from The New York Times

NLU Challenges: Reading Between the Lines

- Lots of understanding is <u>only</u> implied from text.

> "The car stopped about 12 inches above him as he flattened himself at the bottom of the pit."



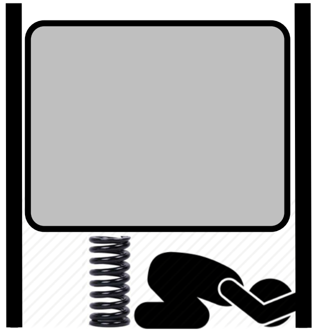
  - *the car is significantly heavier than the man;*
  - *he had nowhere to go at the bottom of the pit;*
  - *if he didn't flatten himself, he would have died;*
  - *...*

The sentence from The New York Times

NLU Challenges: Reading Between the Lines

- Lots of understanding is only implied from text.

"The car stopped about 12 inches above him as he flattened himself at the bottom of the pit."



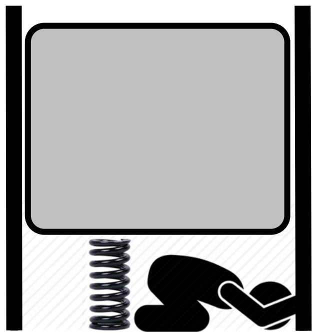
  - *the car is significantly heavier than the man;*
  - *he had nowhere to go at the bottom of the pit;*
  - *if he didn't flatten himself, he would have died;*
  - ...

The sentence from The New York Times

NLU Challenges: Reading Between the Lines

- Lots of understanding is <u>only</u> implied from text.

"The car stopped about 12 inches above him as he flattened himself at the bottom of the pit."



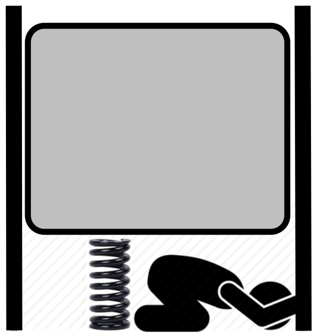
  - *the car is significantly heavier than the man;*
  - *he had nowhere to go at the bottom of the pit;*
  - *if he didn't flatten himself, he would have died;*
  - *...*

The sentence from The New York Times

NLU Challenges: Reading Between the Lines

- Lots of understanding is only implied from text.

"The car stopped about 12 inches above him as he flattened himself at the bottom of the pit."



  - *the car is significantly heavier than the man;*
  - *he had nowhere to go at the bottom of the pit;*
  - *if he didn't flatten himself, he would have died;*
  - *…*

Common-sense understanding

[Aristotle; Avicenna; Descartes; others]

The sentence from The New York Times

NLU Challenges: Small bits, Big conclusions

The story from The New York Times

NLU Challenges: Small bits, Big conclusions

- "Reasoning" as the process of combining facts and beliefs, to make decisions.
  [Johnson-Laird, 1980]

- Why was he crying?
  - Maybe he was scared.
  - Maybe he was injured.

- Many other forms of "reasoning":
  - Inductive, deductive, analogy, quantitative, etc.

The story from The New York Times

NLU Challenges: Small bits, Big conclusions

- "Reasoning" as the process of combining facts and beliefs, to make decisions.

  [Johnson-Laird, 1980]

"… his weeps attracted the attention of a porter"

- Why was he crying?
  - Maybe he was scared.
  - Maybe he was injured.

- Many other forms of "reasoning":
  - Inductive, deductive, analogy, quantitative, etc.

The story from The New York Times

NLU Challenges: Small bits, Big conclusions

- "Reasoning" as the process of combining facts and beliefs, to make decisions.
  [Johnson-Laird, 1980]

"... his weeps attracted the attention of a porter"

- Why was he crying?
  - Maybe he was scared.
  - Maybe he was injured.

- Many other forms of "reasoning":
  - Inductive, deductive, analogy, quantitative, etc.

The story from The New York Times

NLU Challenges: Small bits, Big conclusions

- "Reasoning" as the process of combining facts and beliefs, to make decisions.

  [Johnson-Laird, 1980]

"… his weeps attracted the attention of a porter"

- Why was he crying?
  - Maybe he was scared.
  - Maybe he was injured.

- Many other forms of "reasoning":
  - Inductive, deductive, analogy, quantitative, etc.

The story from The New York Times

NLU Challenges: Small bits, Big conclusions

- "Reasoning" as the process of combining facts and beliefs, to make decisions.

  [Johnson-Laird, 1980]

"… his weeps attracted the attention of a porter"

- Why was he crying?
  - Maybe he was scared.
  - Maybe he was injured.

Abductive reasoning [Peirce, 1883]

- Many other forms of "reasoning":
  - Inductive, deductive, analogy, quantitative, etc.

The story from The New York Times

NLU Challenges: Small bits, Big conclusions

- "Reasoning" as the process of combining facts and beliefs, to make decisions.
  [Johnson-Laird, 1980]

"… his weeps attracted the attention of a porter"

- Why was he crying?
  - Maybe he was scared.
  - Maybe he was injured.

Abductive reasoning [Peirce, 1883]

- Many other forms of "reasoning":
  - Inductive, deductive, analogy, quantitative, etc.

The story from The New York Times

Thesis Structure and Challenges Addressed

Thesis Structure and Challenges Addressed

Approach

Thesis Structure and Challenges Addressed

Approach

❶ System design

Thesis Structure and Challenges Addressed

Approach

❶ System design

❷ Evaluation

Thesis Structure and Challenges Addressed

Approach

❶ System design

❷ Evaluation

❸ Formalism

Thesis Structure and Challenges Addressed

Approach \ Challenge	Variability	Ambiguity	Reasoning	Common-sense understanding
❶ System design				
❷ Evaluation				
❸ Formalism				

Thesis Structure and Challenges Addressed

Challenge / Approach	Variability	Ambiguity	Reasoning	Common-sense understanding
❶ System design			Abductive reasoning	
❷ Evaluation				
❸ Formalism				

Thesis Structure and Challenges Addressed

Approach \ Challenge	Variability	Ambiguity	Reasoning	Common-sense understanding
❶ System design	Semantic similarity Semantic abstraction	Grounding language	Abductive reasoning	
❷ Evaluation				
❸ Formalism				

Thesis Structure and Challenges Addressed

Approach \ Challenge	Variability	Ambiguity	Reasoning	Common-sense understanding
❶ System design	Semantic similarity Semantic abstraction	Grounding language	Abductive reasoning	
❷ Evaluation			Reasoning with multiple pieces of text	
❸ Formalism				

Thesis Structure and Challenges Addressed

Challenge / Approach	Variability	Ambiguity	Reasoning	Common-sense understanding
❶ System design	Semantic similarity Semantic abstraction	Grounding language	Abductive reasoning	
❷ Evaluation			Reasoning with multiple pieces of text	Reasoning with common-sense understanding
❸ Formalism				

Thesis Structure and Challenges Addressed

<table>
<tr><th>Approach \ Challenge</th><th>Variability</th><th>Ambiguity</th><th>Reasoning</th><th>Common-sense understanding</th></tr>
<tr><td>❶ System design</td><td>Semantic similarity
Semantic abstraction</td><td>Grounding language</td><td>Abductive reasoning</td><td></td></tr>
<tr><td>❷ Evaluation</td><td></td><td></td><td>Reasoning with multiple pieces of text</td><td>Reasoning with common-sense understanding</td></tr>
<tr><td>❸ Formalism</td><td colspan="3">A framework to study reasoning,
in presence of variability and ambiguity.</td><td></td></tr>
</table>

Thesis Statement

- Progress in automated question answering could be facilitated by incorporating the ability to reason over natural language abstractions and world knowledge.

- More challenging, yet realistic QA datasets pose problems to current technologies; hence, more opportunities for improvement.

Road map

- Introduction and motivation
- Part 1: Reasoning-Driven System Design
    - QA as Subgraph Optimization on Tabular Knowledge *[IJCAI'16]*
    - QA with Semantic Abstractions of Raw Text *[AAAI'18]*
    - Learning to Pay Attention to Essential Terms in Questions *[CoNLL'17]*
- Part 2: Moving the Peaks Higher: More Challenging Datasets
    - A QA Benchmark for Reasoning on Multiple Sentences *[NAACL'18]*
    - A QA Benchmark for Temporal Common-sense *[Submitted]*
- Part 3: Formal Study of Reasoning in Natural Language
    - Capabilities and Limitations of Reasoning in Natural Language *[In submission]*
- Conclusion

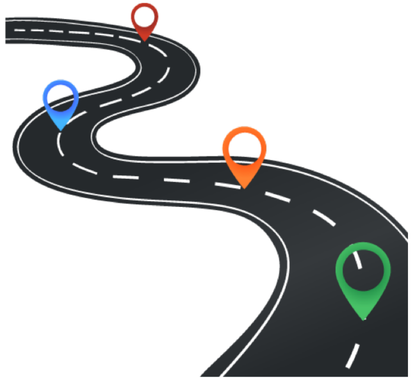
Road map

- Introduction and motivation
- Part 1: Reasoning-Driven System Design
  - QA as Subgraph Optimization on Tabular Knowledge *[IJCAI'16]*
  - QA with Semantic Abstractions of Raw Text *[AAAI'18]*
  - Learning to Pay Attention to Essential Terms in Questions *[CoNLL'17]*
- Part 2: Moving the Peaks Higher: More Challenging Datasets
  - A QA Benchmark for Reasoning on Multiple Sentences *[NAACL'18]*
  - A QA Benchmark for Temporal Common-sense *[Submitted]*
- Part 3: Formal Study of Reasoning in Natural Language
  - Capabilities and Limitations of Reasoning in Natural Language *[In submission]*
- Conclusion

before
Prelim
after

Road Map

- **Part 1:** Reasoning-Driven System Design
  - QA as Subgraph Optimization on Tabular Knowledge *[IJCAI'16]*
    - Motivation
    - Knowledge as Tables
    - Reasoning on Knowledge
    - Experimental results



QA as Subgraph Optimization on Tabular Internal Knowledge: Overview

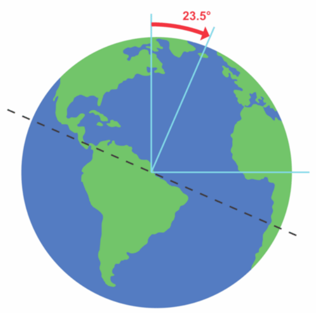
[Clark et al, 2015]

QA as Subgraph Optimization on Tabular Internal Knowledge: Overview

[Clark et al, 2015]

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

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

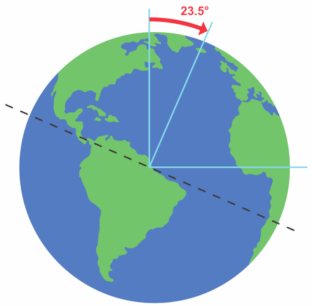


QA as Subgraph Optimization on Tabular Internal Knowledge: Overview

- Standardized science exams. [Clark et al, 2015]
- Simple language; machines require the ability use the knowledge and abstract over it.
- The "knowledge" encoded within the solver.

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

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

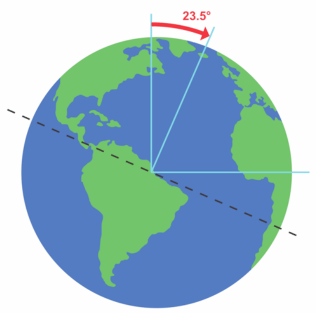


QA as Subgraph Optimization on Tabular Internal Knowledge: Overview

- Standardized science exams. [Clark et al, 2015]
- Simple language; machines require the ability use the knowledge and abstract over it.
- The "knowledge" encoded within the solver.

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

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



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

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

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!*

New Zealand

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!*

New Zealand shortest

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!*

New Zealand shortest night

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

(A) June

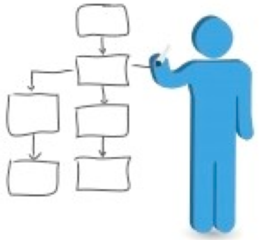
(B) March

(C) December

(D) September

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

Semi-Structured Inference



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

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

Semi-Structured Inference

Question: 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, ….*
  - Reasoning: putting together pieces of knowledge in a principled way.
  - Goal: overcome brittleness

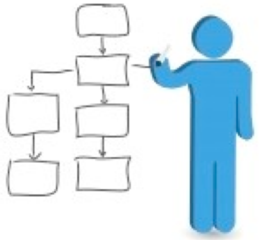
✓ principled approach, explainable answers

✓ robust to variations

Semi-Structured Inference

Question: 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, ….
  - Reasoning: putting together pieces of knowledge in a principled way.
  - Goal: overcome brittleness

✓ principled approach, explainable answers

✓ robust to variations

Semi-Structured Inference

Question: 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, ….
  - Reasoning: putting together pieces of knowledge in a principled way.
  - Goal: overcome brittleness

New York

Longest Day

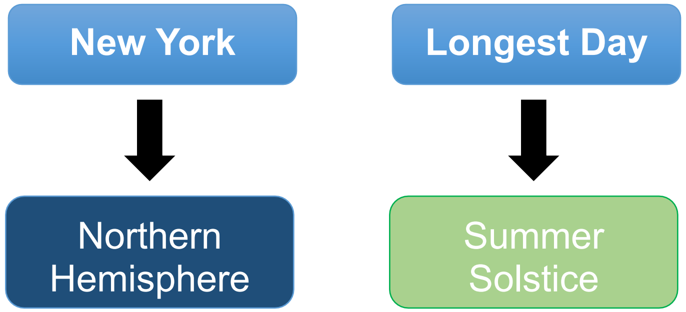
✓ principled approach, explainable answers

✓ robust to variations

Semi-Structured Inference

> ***Question:*** 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, ….
  - Reasoning: putting together pieces of knowledge in a principled way.
  - Goal: overcome brittleness



✓ principled approach, explainable answers

✓ robust to variations

Semi-Structured Inference

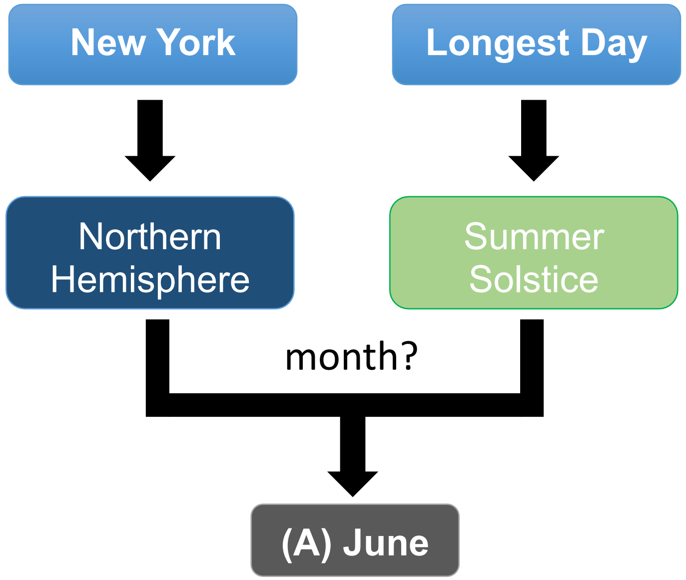
Question: In New York State, the longest period of daylight occurs during which month?

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

- Structured, Multi-Step Reasoning
  - Science knowledge in small, manageable, swappable pieces:
    regions, hemispheres, solstice, ….
  - Reasoning: putting together pieces of knowledge in a principled way.
  - Goal: overcome brittleness

✓ principled approach, explainable answers

✓ robust to variations



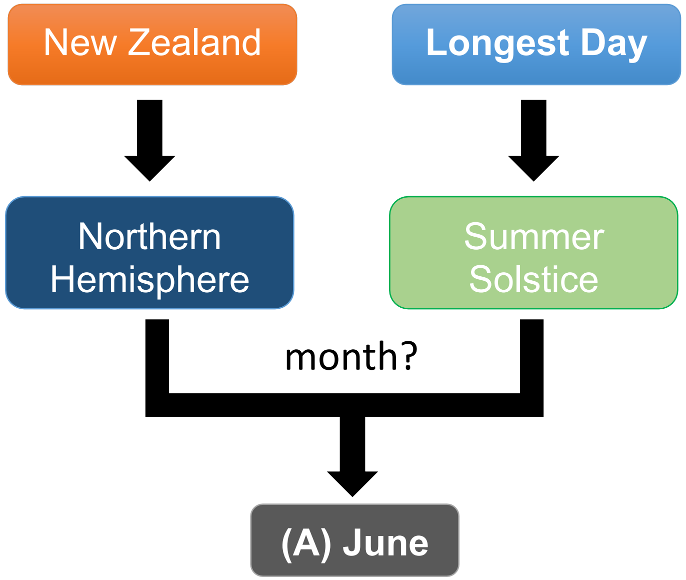
Semi-Structured Inference

Question: 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, ….
  - Reasoning: putting together pieces of knowledge in a principled way.
  - Goal: overcome brittleness

✓ principled approach, explainable answers

✓ robust to variations



Semi-Structured Inference

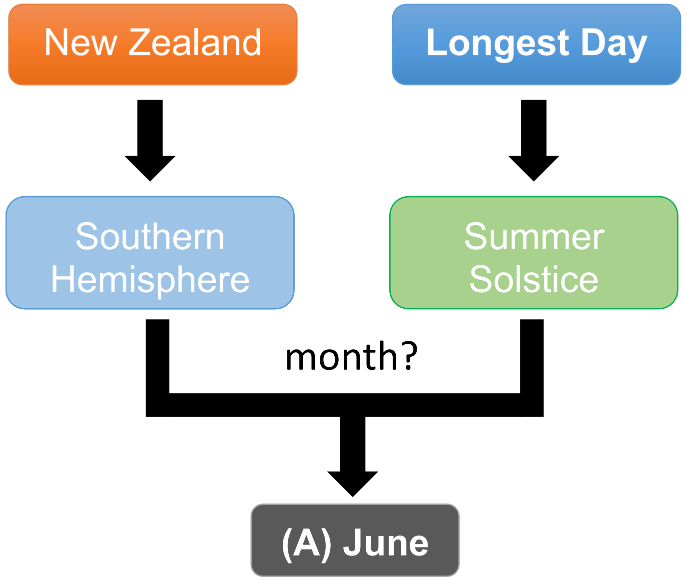
Question: In New York State, the longest period of daylight occurs during which month?

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

- Structured, Multi-Step Reasoning
  - Science knowledge in small, manageable, swappable pieces:
    regions, hemispheres, solstice, ….
  - Reasoning: putting together pieces of knowledge in a principled way.
  - Goal: overcome brittleness

✓ principled approach, explainable answers

✓ robust to variations



Semi-Structured Inference

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

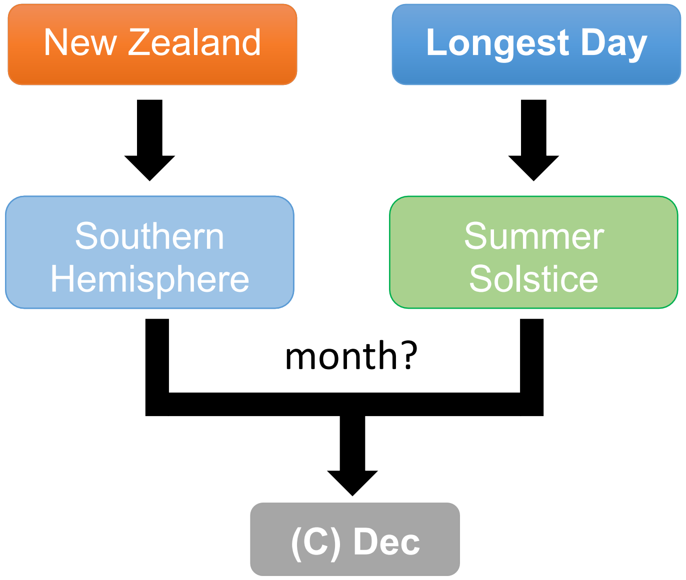
(A) June (B) March (C) December (D) September

- Structured, Multi-Step Reasoning
  - Science knowledge in small, manageable, swappable pieces:

    regions, hemispheres, solstice, ….
  - Reasoning: putting together pieces of knowledge in a principled way.
  - Goal: overcome brittleness

✓ principled approach, explainable answers

✓ robust to variations



Semi-Structured Inference

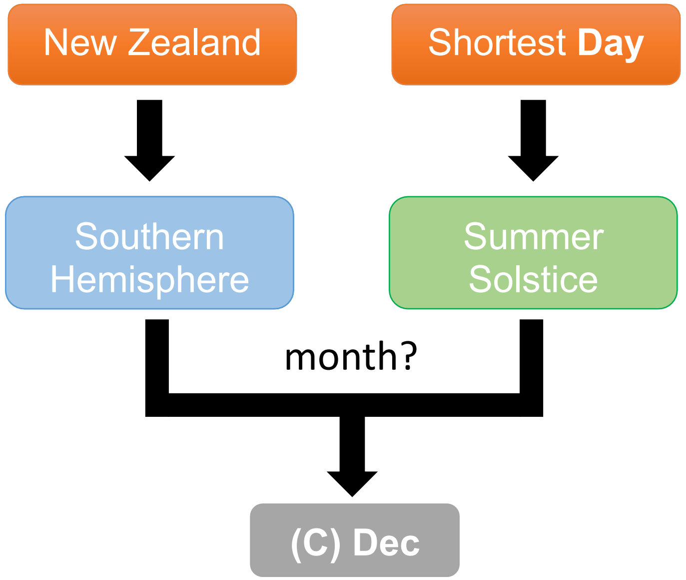
Question: In New York State, the longest period of daylight occurs during which month?

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

- Structured, Multi-Step Reasoning
  - Science knowledge in small, manageable, swappable pieces:
    regions, hemispheres, solstice, ….
  - Reasoning: putting together pieces of knowledge in a principled way.
  - Goal: overcome brittleness

✓ principled approach, explainable answers

✓ robust to variations



Semi-Structured Inference

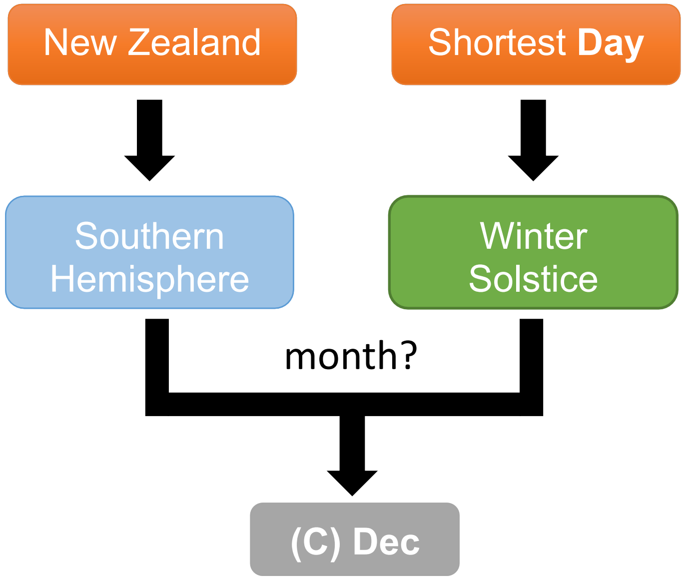
Question: In New York State, the longest period of daylight occurs during which month?

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

- Structured, Multi-Step Reasoning
  - Science knowledge in small, manageable, swappable pieces:
    regions, hemispheres, solstice, ….
  - Reasoning: putting together pieces of knowledge in a principled way.
  - Goal: overcome brittleness

✓ principled approach, explainable answers

✓ robust to variations



Semi-Structured Inference

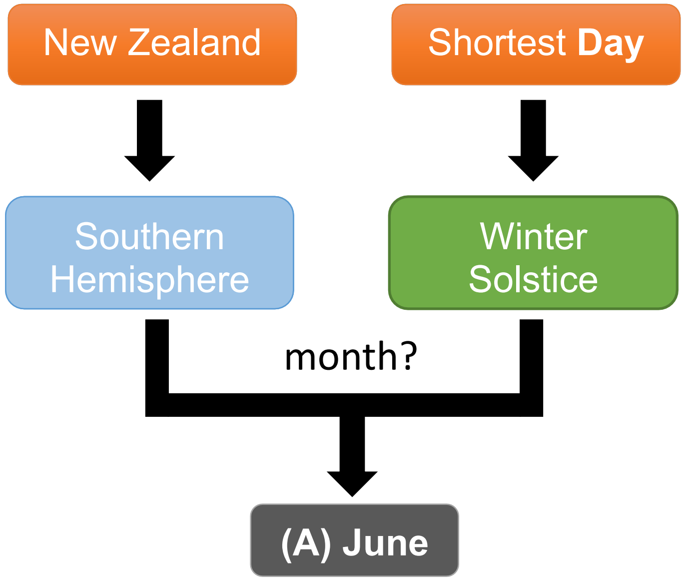
Question: In New York State, the longest period of daylight occurs during which month?

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

- Structured, Multi-Step Reasoning
  - Science knowledge in small, manageable, swappable pieces:
    regions, hemispheres, solstice, ….
  - Reasoning: putting together pieces of knowledge in a principled way.
  - Goal: overcome brittleness

✓ principled approach, explainable answers

✓ robust to variations



Semi-Structured Inference

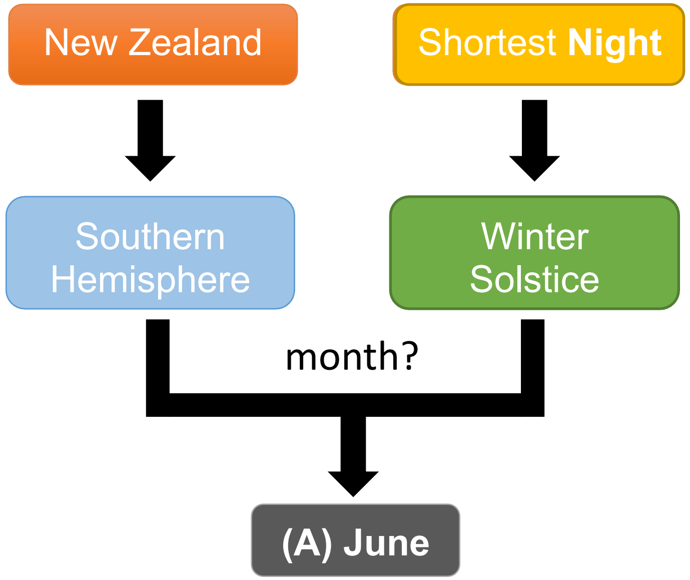
Question: In New York State, the longest period of daylight occurs during which month?

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

- Structured, Multi-Step Reasoning
  - Science knowledge in small, manageable, swappable pieces:
    regions, hemispheres, solstice, ….
  - Reasoning: putting together pieces of knowledge in a principled way.
  - Goal: overcome brittleness

✓ principled approach, explainable answers

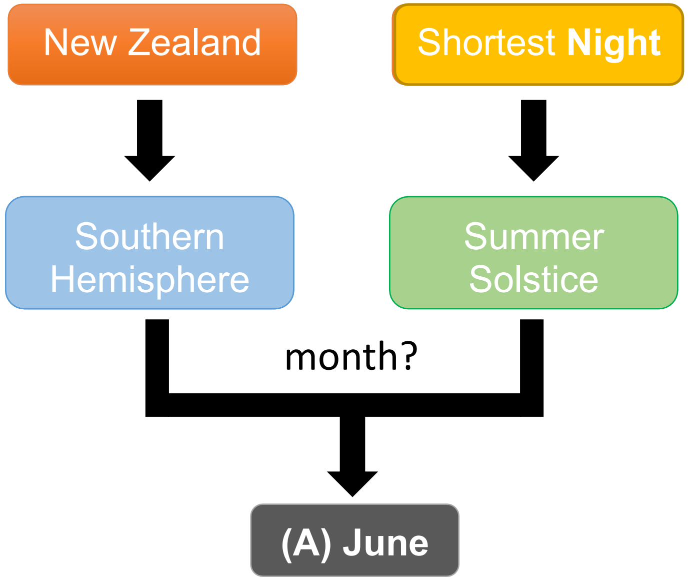
✓ robust to variations



Semi-Structured Inference

Question: 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, ….
  - Reasoning: putting together pieces of knowledge in a principled way.
  - Goal: overcome brittleness



✓ principled approach, explainable answers

✓ robust to variations

Semi-Structured Inference

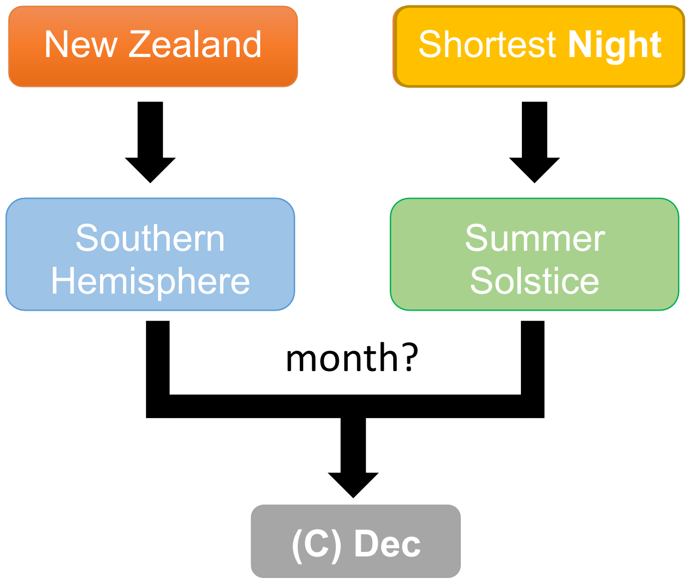
Question: In New York State, the longest period of daylight occurs during which month?

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

- Structured, Multi-Step Reasoning
  - Science knowledge in small, manageable, swappable pieces:
    regions, hemispheres, solstice, ….
  - Reasoning: putting together pieces of knowledge in a principled way.
  - Goal: overcome brittleness

✓ principled approach, explainable answers

✓ robust to variations



Semi-Structured Inference

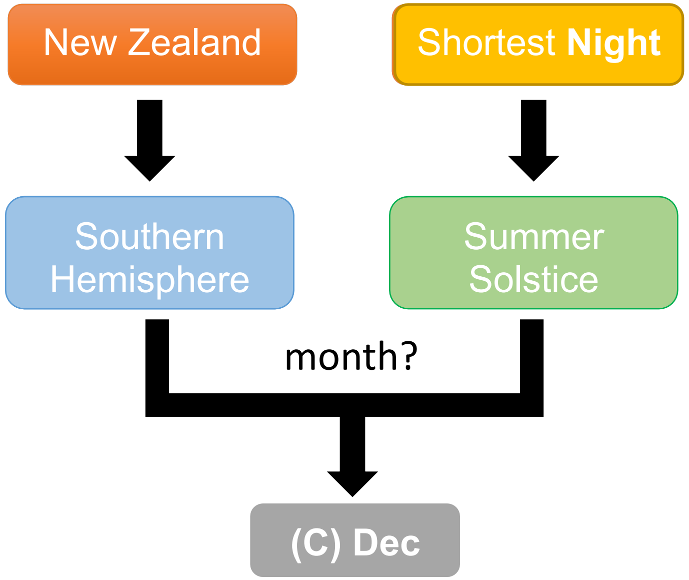
Question: In New York State, the longest period of daylight occurs during which month?

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

- Structured, Multi-Step Reasoning
  - Science knowledge in small, manageable, swappable pieces:
    regions, hemispheres, solstice, ….
  - Reasoning: putting together pieces of knowledge in a principled way.
  - Goal: overcome brittleness

✓ principled approach, explainable answers

✓ robust to variations



How can we achieve this?

Semi-Structured Inference: High-level View

Question Answering

as Global Reasoning

over Semi-Structured Knowledge

Question Answering

Semi-Structured Inference: High-level View

Question Answering

as **Global Reasoning**

over **Semi-Structured Knowledge**

Question Answering

Knowledge Representation

Structured,

yet flexible

Semi-Structured Inference: High-level View

Question Answering

as Global Reasoning

over Semi-Structured Knowledge

Question Answering

Knowledge Representation

Structured,

yet flexible

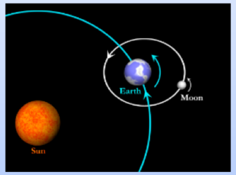
Reasoning

Effective, scalable

The Necessary Knowledge

The Knowledge Atlas: 12 key sections

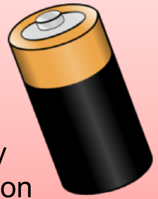
Celestial Phenomena
sun
moon
stars
day/night,
rotation
revolution



The Earth
air
water
land
weather
precipitation
erosion

Matter
solid/liquid/gas
properties
conductivity
texture
temperature
measuring tools

Energy
forms
energy transfer
heat
electricity
chemical energy
energy conversion



Forces
gravity
magnetism
force
friction
pull/pushing
attraction

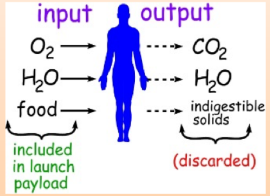
Living things
living
nonliving
characteristics
animals
plants
fish

Inheritance
inherited traits
resemblance
acquired traits
learned traits
body features
skills

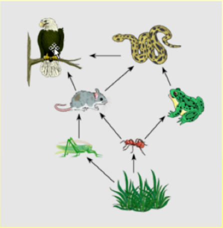
The Environment and Adaptation
senses
habitats
behavior
camouflage
survival

Continuity of Life
life cycle
life span
offspring
reproduction
coloration
mating

Life Functions
breathing
growing
eating
food
air
water



Interdependence
food web
producers
consumers
decomposers
predators
prey



Human Impact
human activities
environment
ecosystem
pollution
conservation
deforestation

Knowledge as Frames

Frame Semantics
[Minsky, 1974; Fillmore, 1977]

Knowledge as Frames

Orbital events

Hemisphere: ?

Orbital events: ?

Month: ?

Frame Semantics

[Minsky, 1974; Fillmore, 1977]

Knowledge as Frames

Orbital events

Hemisphere: ?

Orbital events: ?

Month: ?

Frame Semantics

[Minsky, 1974; Fillmore, 1977]

Knowledge as Frames

Orbital events
Hemisphere: ?
Orbital events: ?
Month: ?

Frame Semantics
[Minsky, 1974; Fillmore, 1977]

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, …

collections of recurring, related, science concepts

Simple structure, flexible content
- Can acquire knowledge in automated and semi-automated ways [Dalvi et al, 2016]

Knowledge as Frames

Unstructured — Structured

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

easy to acquire, difficult to reason with

Frames with free form text

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

“easy” to reason with, difficult to acquire

Orbital events

Hemisphere: ?

Orbital events: ?

Month: ?

Frame Semantics

[Minsky, 1974; Fillmore, 1977]

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, ...

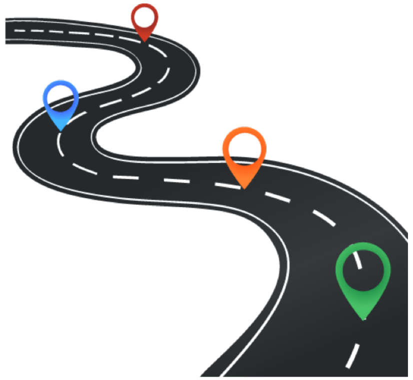
collections of recurring, related, science concepts

Simple structure, flexible content

- Can acquire knowledge in automated and semi-automated ways [Dalvi et al, 2016]

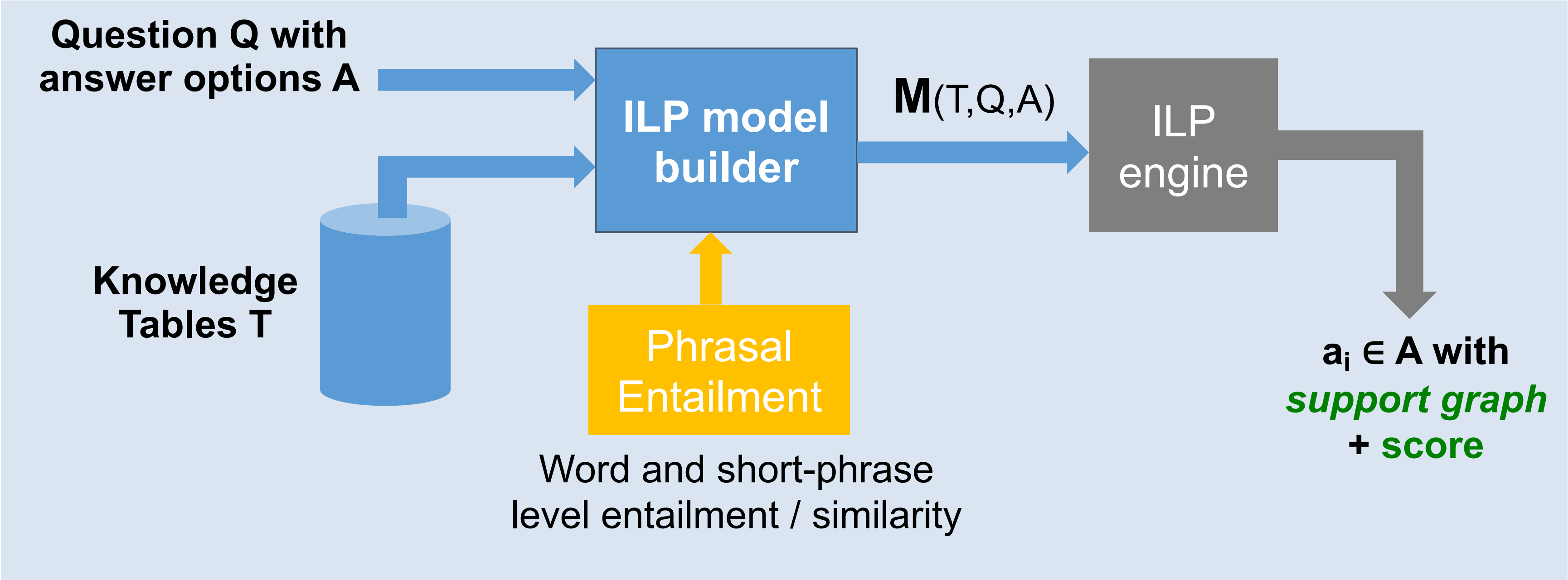
Road Map

- **Part 1:** Reasoning-Driven System Design
  - QA as Subgraph Optimization on Tabular Knowledge *[IJCAI'16]*
    - Motivation
    - Knowledge as Frames
    - Reasoning on Knowledge
    - Experimental results



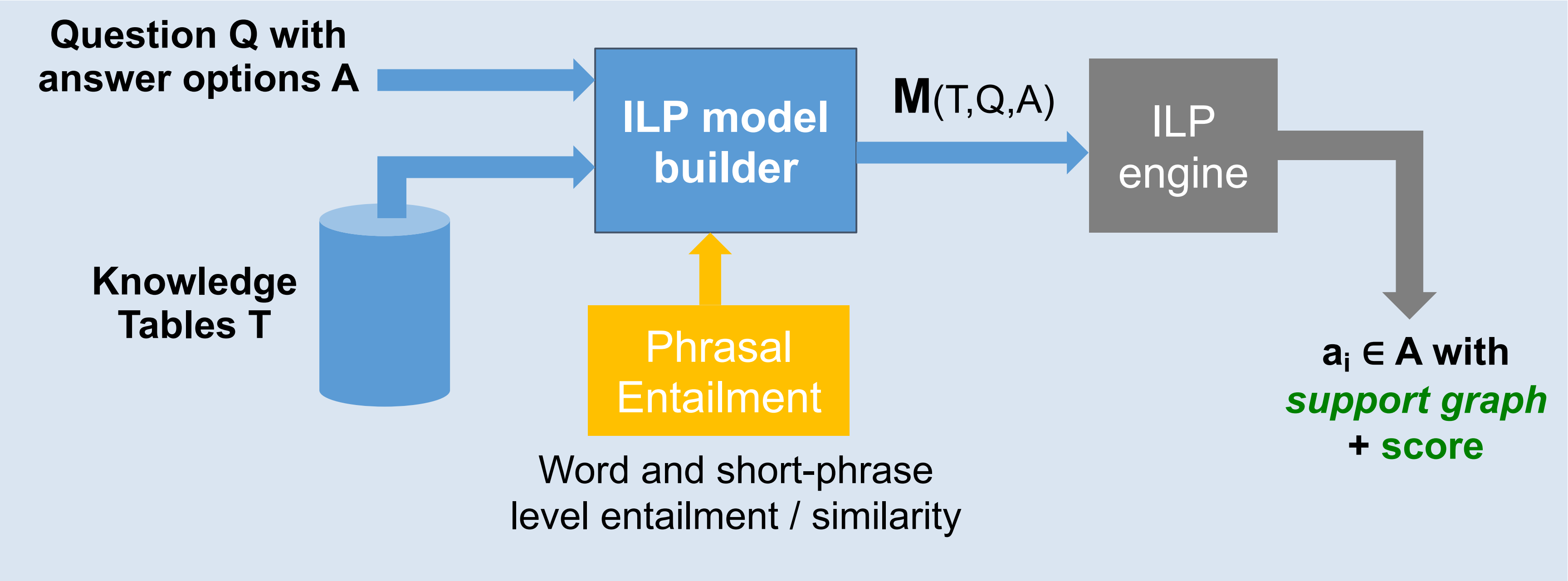
TableILP Solver: An Overview

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



TableILP Solver: An Overview

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



M(T,Q,A) ⟶ $\max \sum_i c_i x_i$ $\quad \forall\, x_i \in \mathbb{N} \cup \{0\}$

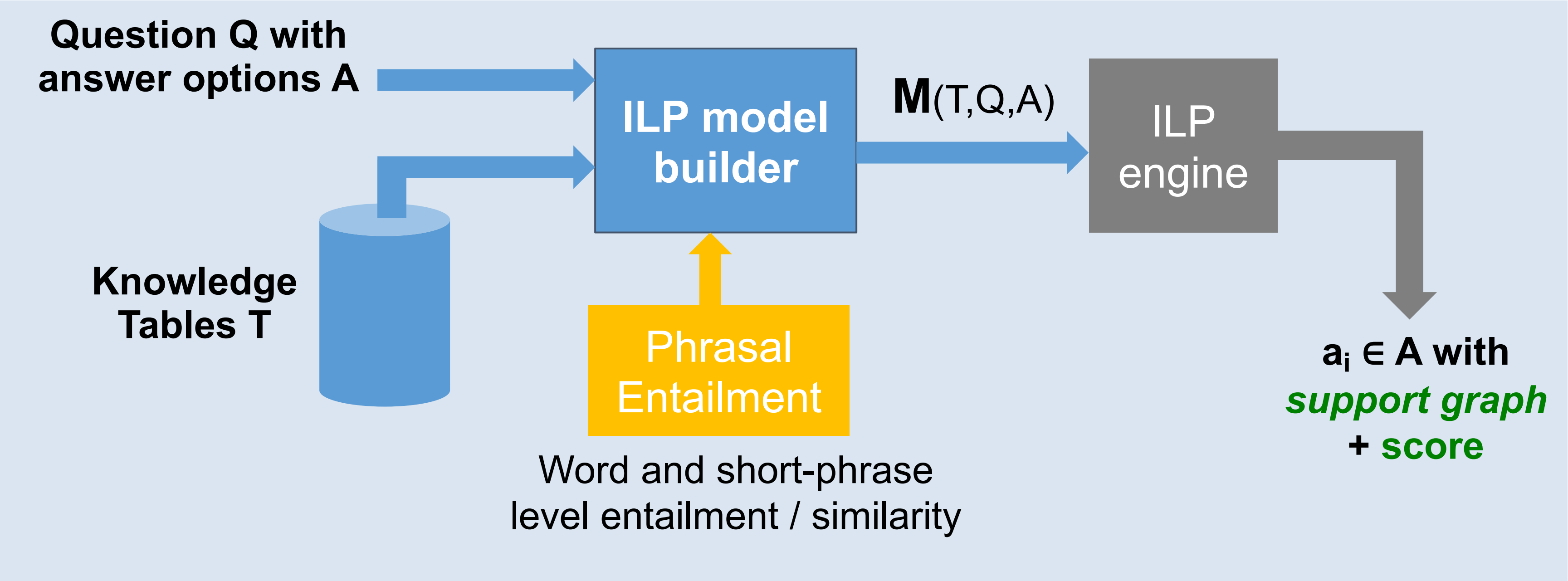
$$\begin{cases} \sum_i a_{1i} x_i \leq b_1 \\ \quad \ldots \\ \sum_i a_{ki} x_i \leq b_k \end{cases}$$

Optimization using Integer Linear Program (**ILP**) formalism

TableILP Solver: An Overview

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

Sources of knowledge:



M(T,Q,A) ⟹

$$\max \sum_i c_i x_i \qquad \forall\, x_i \in \mathbb{N} \cup \{0\}$$

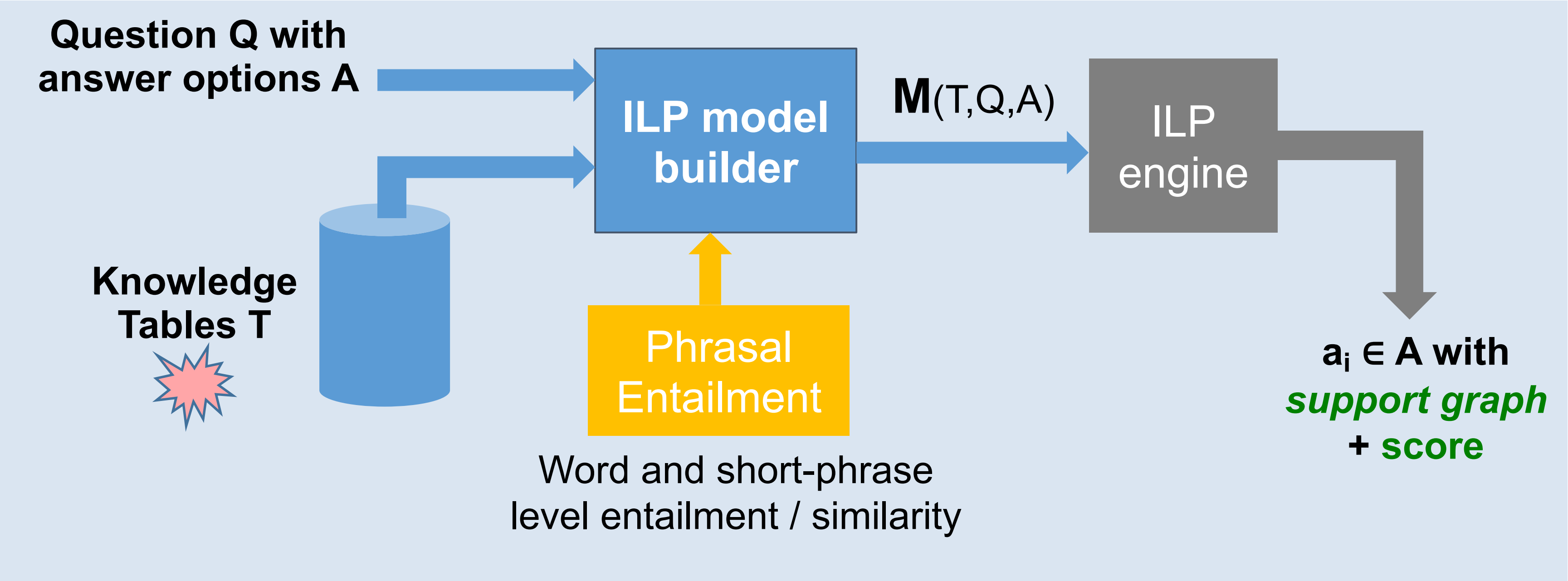
$$\begin{cases} \sum_i a_{1i} x_i \le b_1 \\ \dots \\ \sum_i a_{ki} x_i \le b_k \end{cases}$$

Optimization using Integer Linear Program (**ILP**) formalism

TableILP Solver: An Overview

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

Sources of knowledge:



M(T,Q,A) ⟶

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

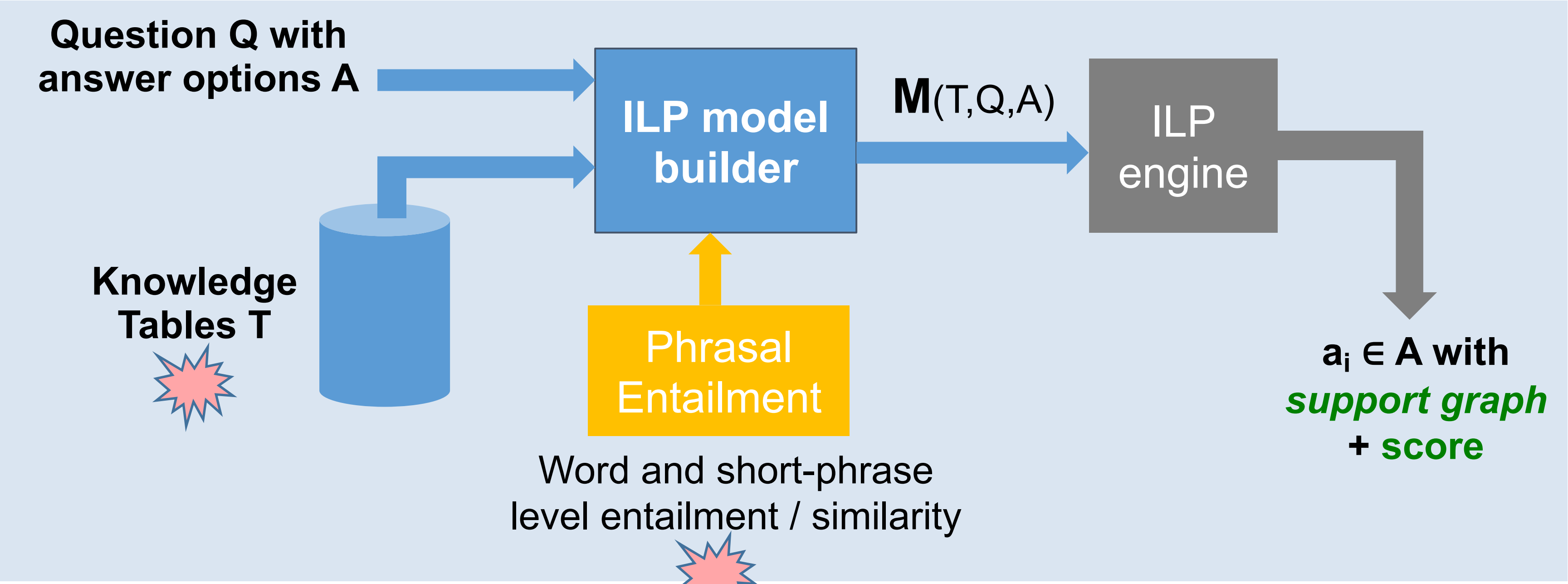
$$\forall\, x_i \in \mathbb{N} \cup \{0\}$$

Optimization using Integer Linear Program (**ILP**) formalism

TableILP Solver: An Overview

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

Sources of knowledge:



M(T,Q,A) ⟶

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

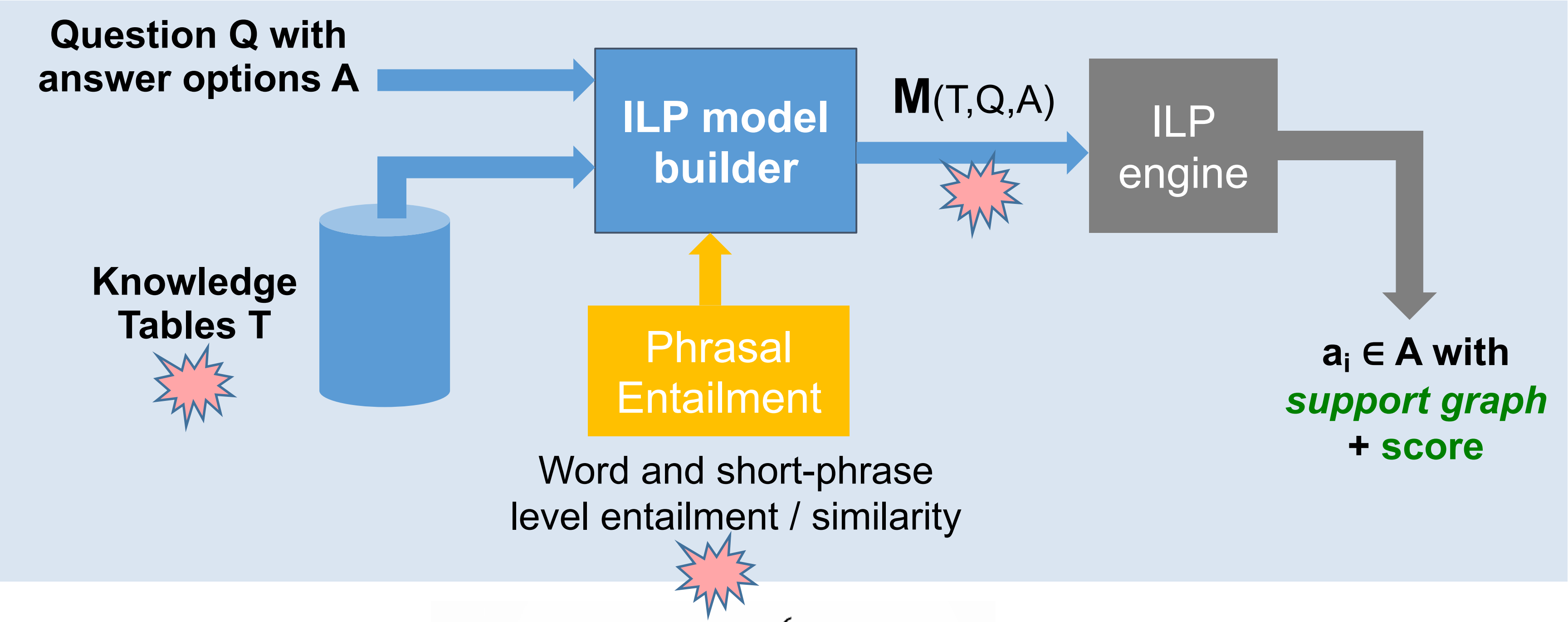
$$\forall\, x_i \in \mathbb{N} \cup \{0\}$$

Optimization using Integer Linear Program (**ILP**) formalism

TableILP Solver: An Overview

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

Sources of knowledge:



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 Program (**ILP**) formalism

TableILP: Main Idea

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

TableILP: Main Idea

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

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

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

TableILP: Main Idea

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

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

How is relevant information expressed in my KB?

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

TableILP: Main Idea

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

Cities, States, Countries

Orbital Events:

Geographical properties & Timing

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

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

TableILP: Main Idea

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

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

TableILP: Main Idea

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

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

TableILP: Main Idea

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
….	….	…

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

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

Abductive reasoning

[Peirce, 1883]

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

TableILP: Main Idea

An ideal Support Graph

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

Link this information to identify the best supported answer!

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

Approach: ILP model

Not so straightforward!

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

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

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

- Many possible "proof structures"
  - single/multi-table, single/multi-row, answer in table header, answer spanning multiple cells
- Must balance reward for connections with penalty for spurious links
- Imperfect lexical "similarity" blackbox
- Partial or missing knowledge in tables
- Question logic (negation, conjunction, comparison)
- Scalability of ILP solvers

ILP model

Operates on lexical units of alignment

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

$$\max \sum_i c_i x_i \qquad \forall\, x_i \in \mathbb{N} \cup \{0\}$$

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

~50 high level constraints + preferences

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

- Which nodes + edges between lexical units are <u>active</u>?

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

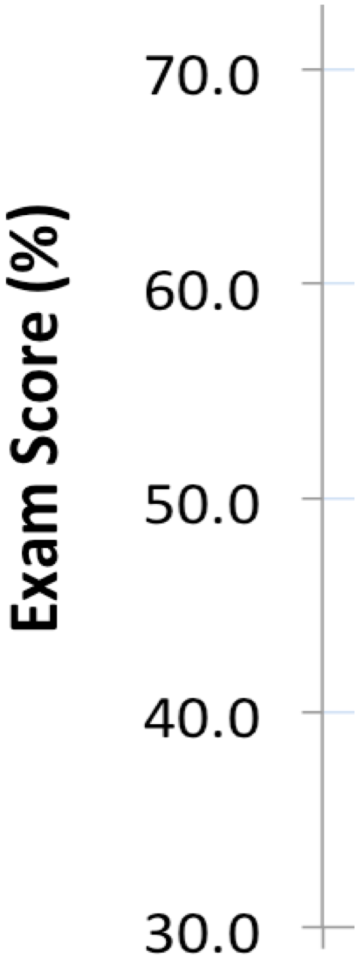
- Reward active units, high lexical match links, column header match, ...
- WH-term boost ("*which form of energy...*"), science-term boost ("*evaporation*")
- Penalize spurious overuse of frequently occurring terms

ILP Model: Constraints

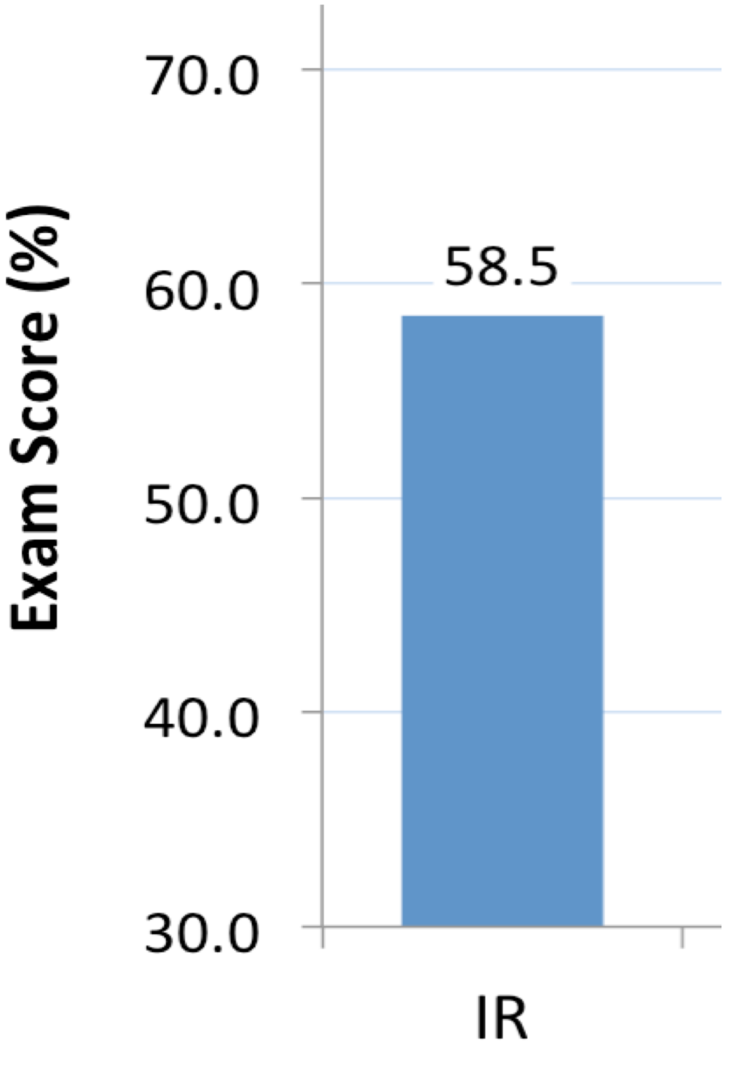
Dual goal: scalability, consider only meaningful support graphs

- **Structural Constraints**
  - Meaningful proof structures
    - connectedness, question coverage, appropriate table use
    - single/multi-table, single/multi-row, etc
  - Simplicity appropriate for 4th / 8th grade
- **Semantic Constraints**
  - Chaining => table joins between semantically similar column pairs
  - Relation matching (ruler measures length, change from water to liquid)

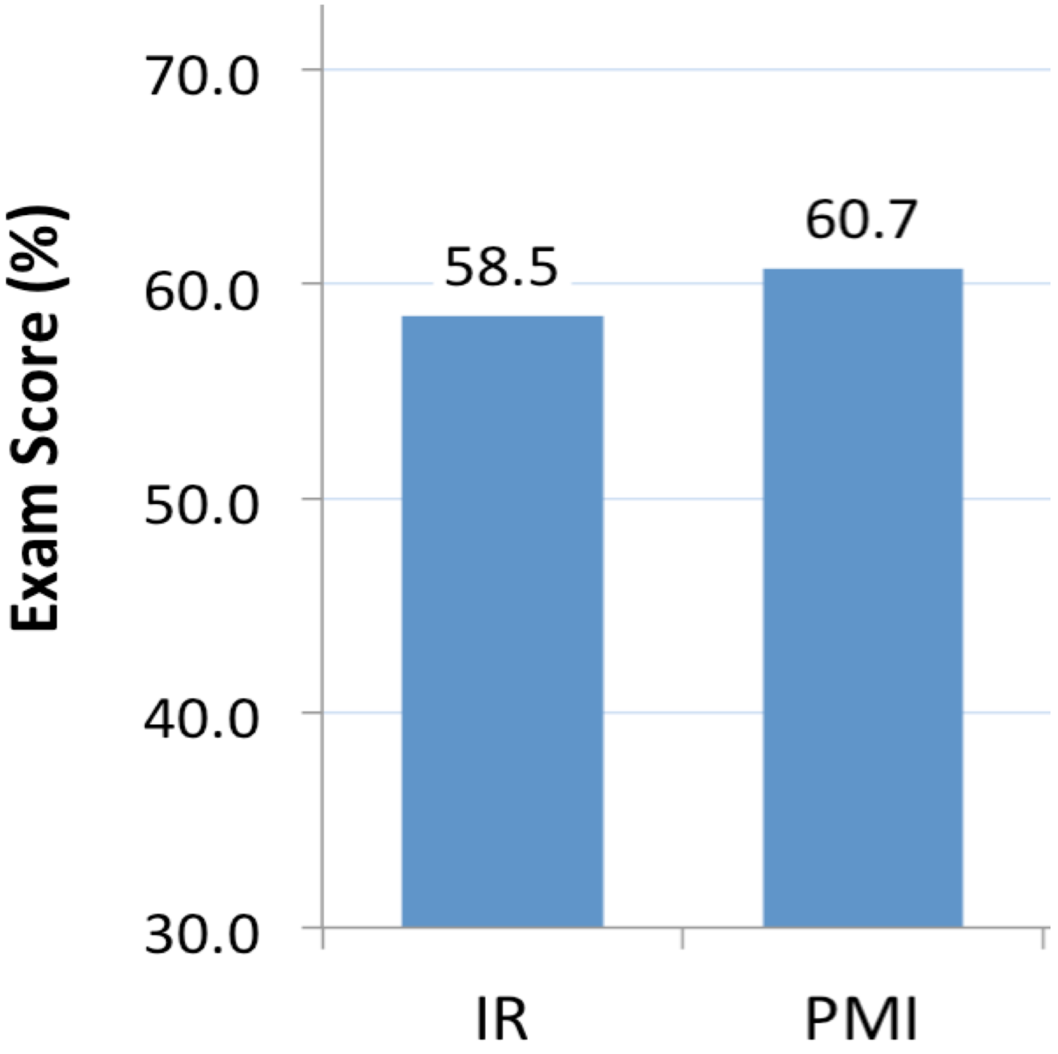
Experimental Results [KKSR’16]



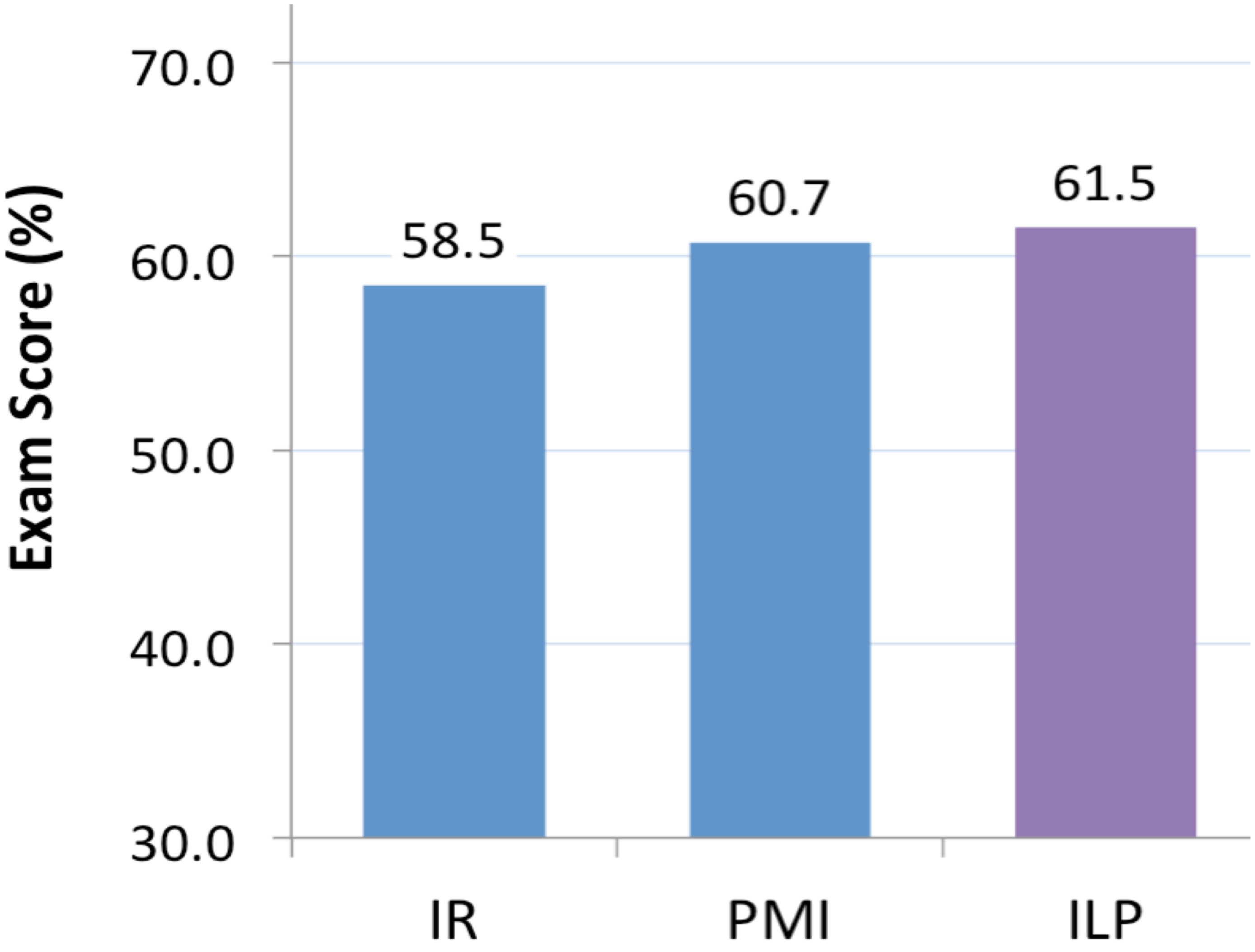
Experimental Results [KKSR'16]



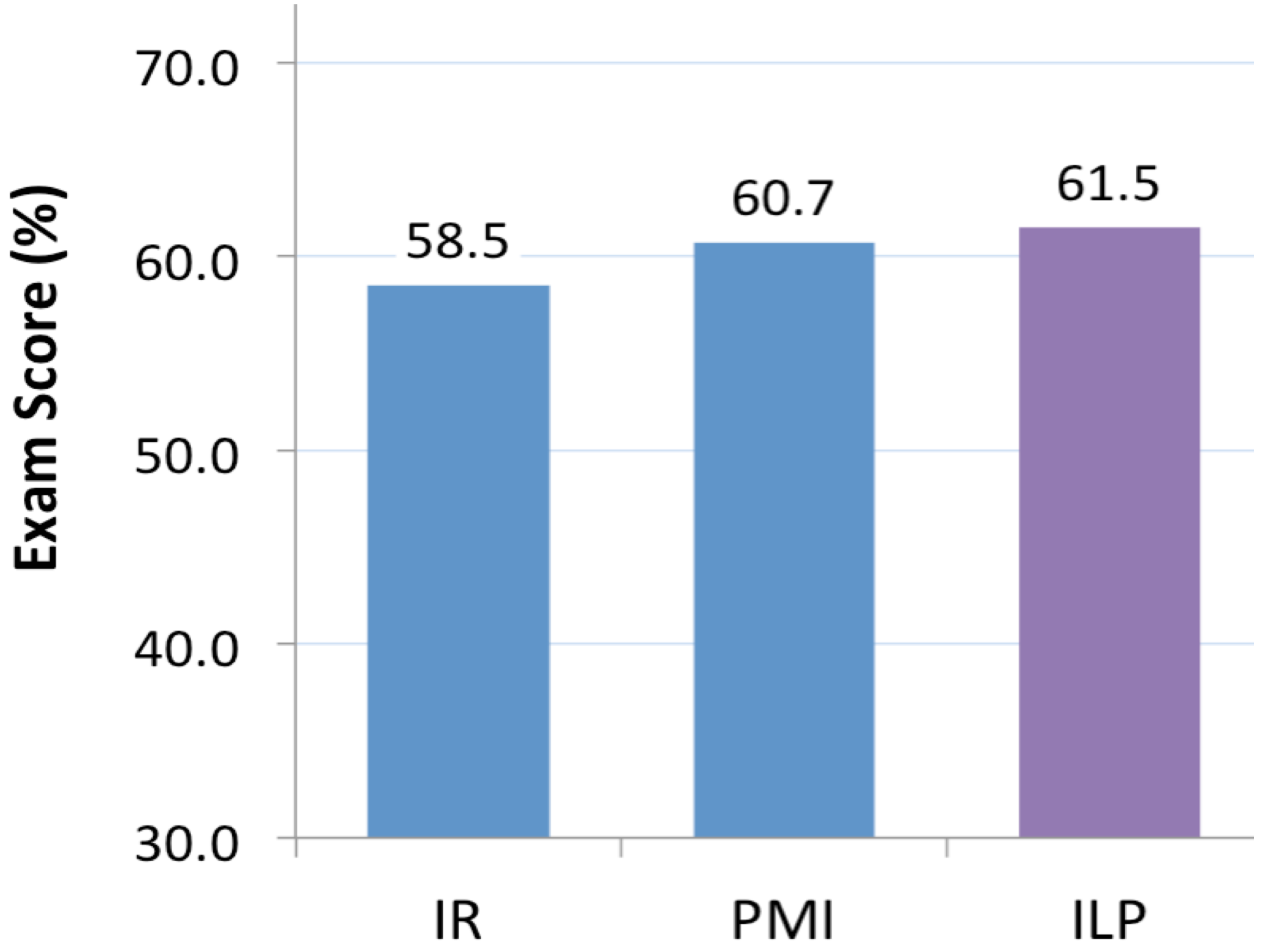
Experimental Results [KKSR'16]



Experimental Results [KKSR'16]



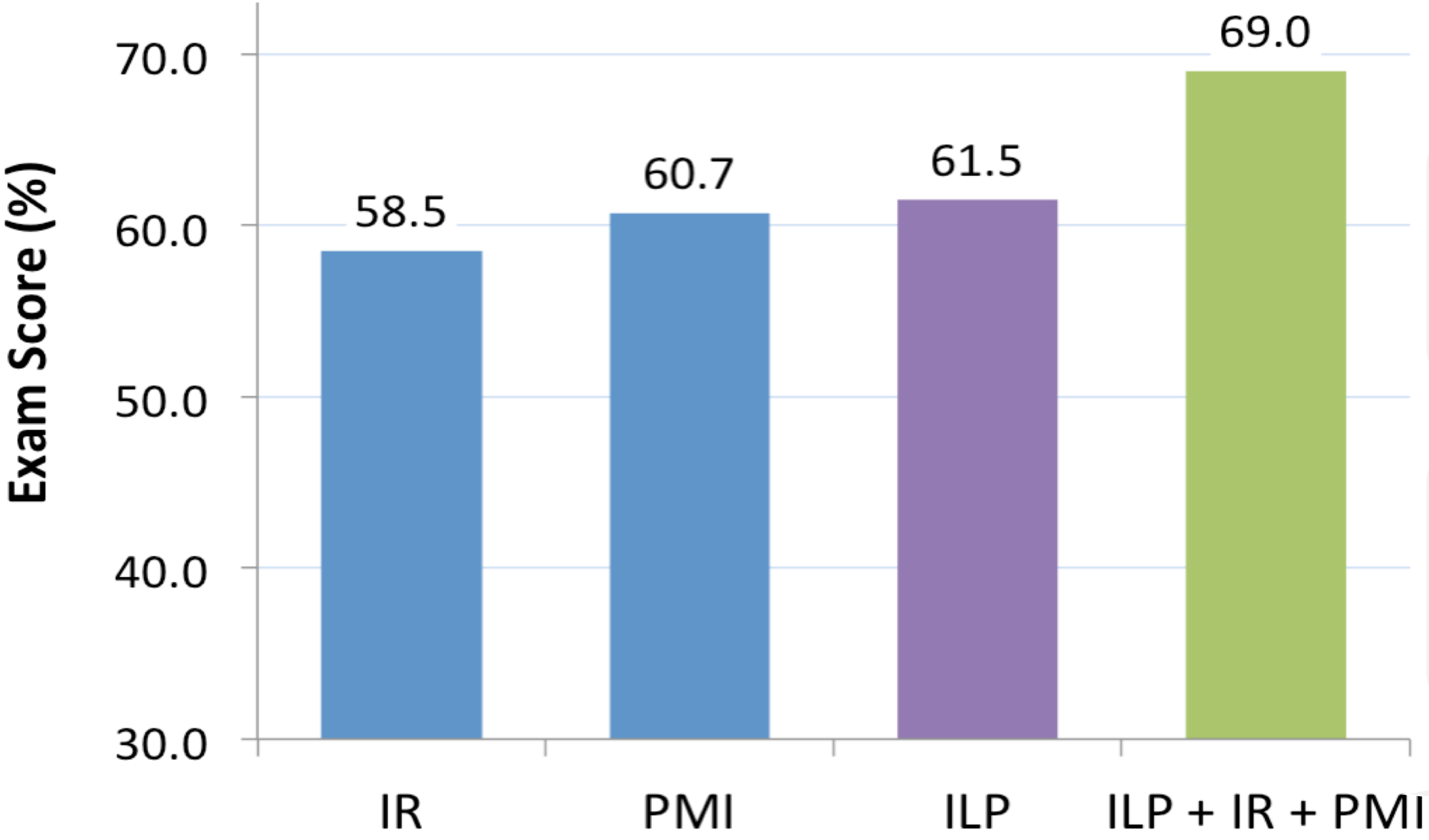
Experimental Results [KKSR'16]



TableILP scores 54% on "IR-hard" questions

(random guessing = 25%)

Experimental Results [KKSR'16]

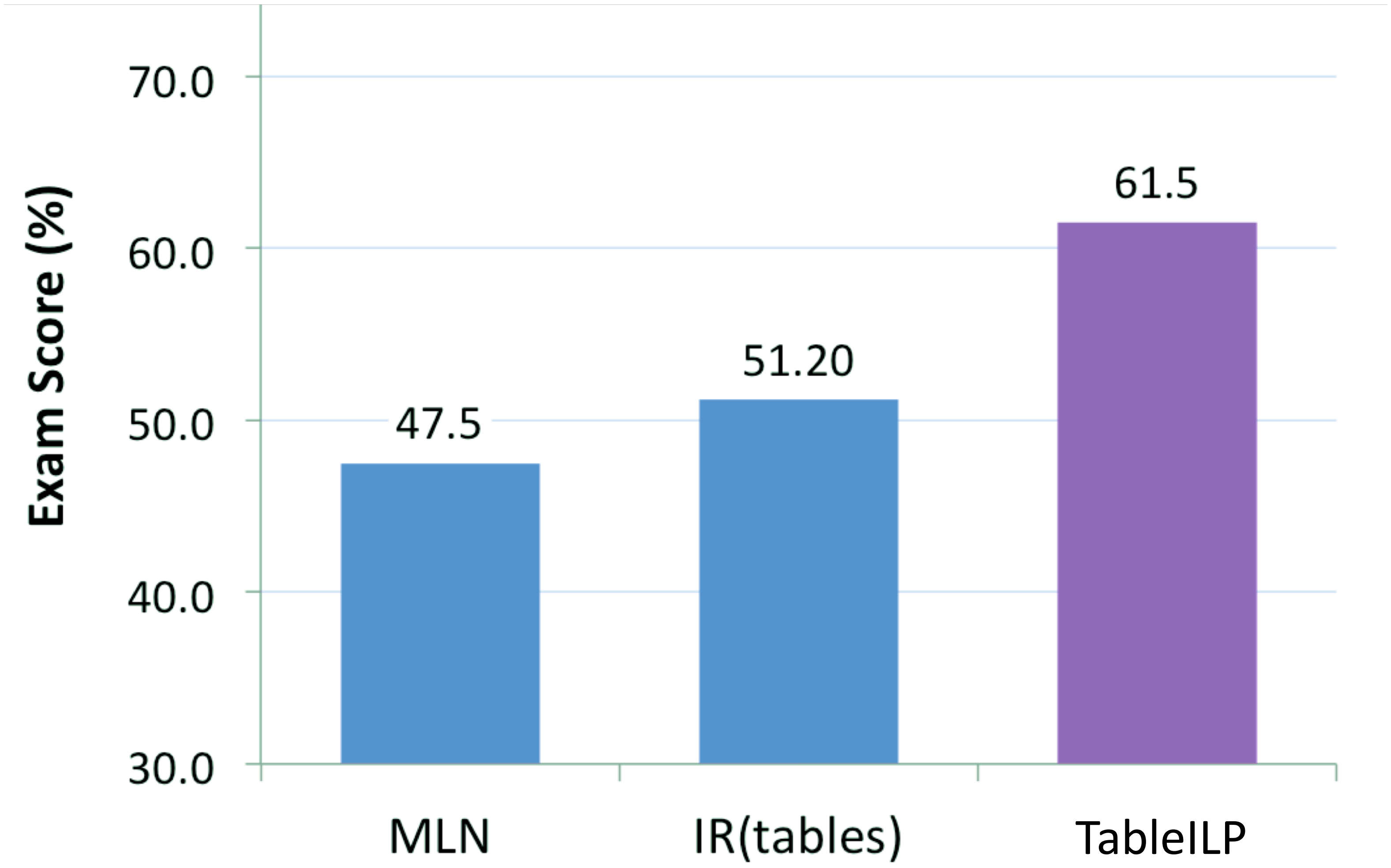


TableILP scores 54% on "IR-hard" questions
(random guessing = 25%)

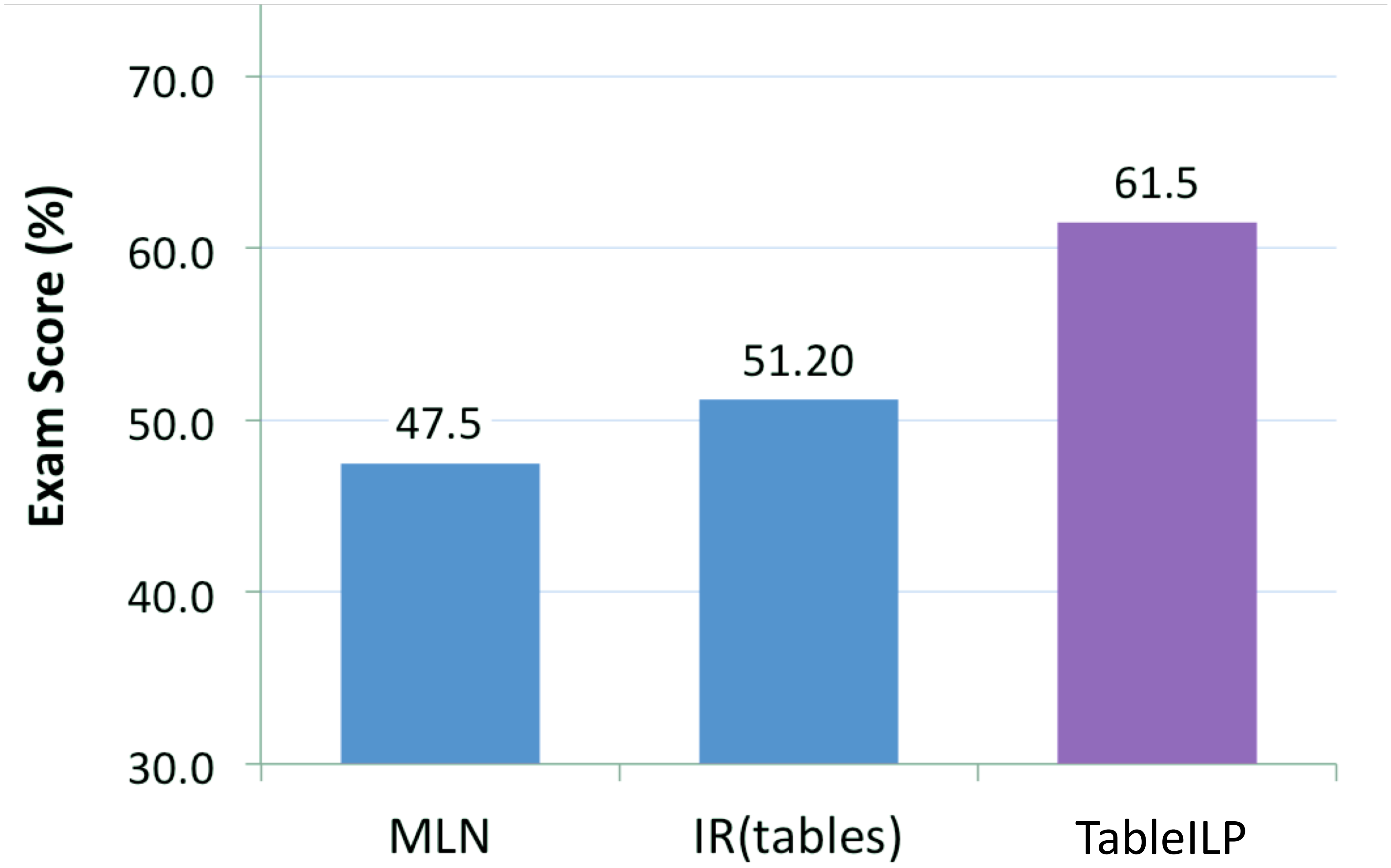
Ensemble performs 8–10% higher than IR baselines

[CEKSTTK'16]

Experimental Results [KKSR'16]



Experimental Results [KKSR'16]



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

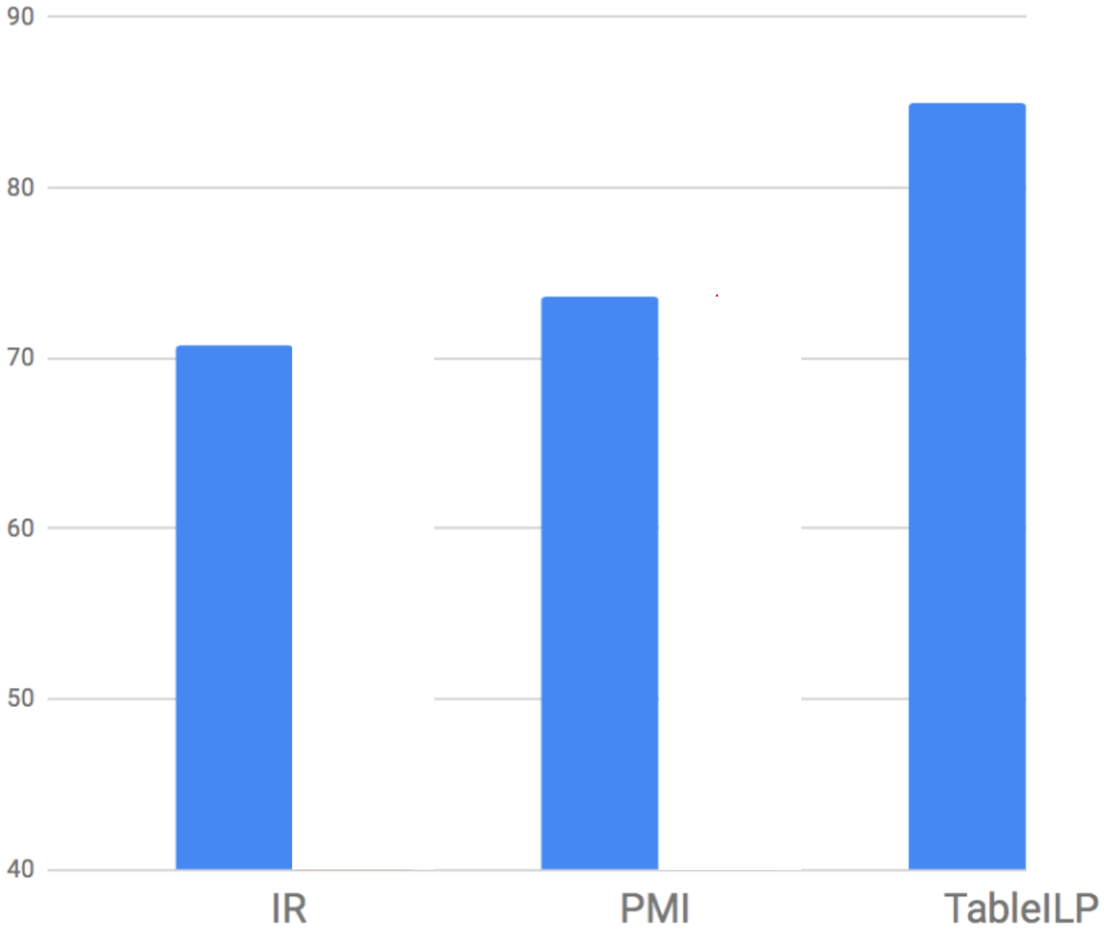
Assessing Brittleness: Question Perturbation

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*

Assessing Brittleness: Question Perturbation

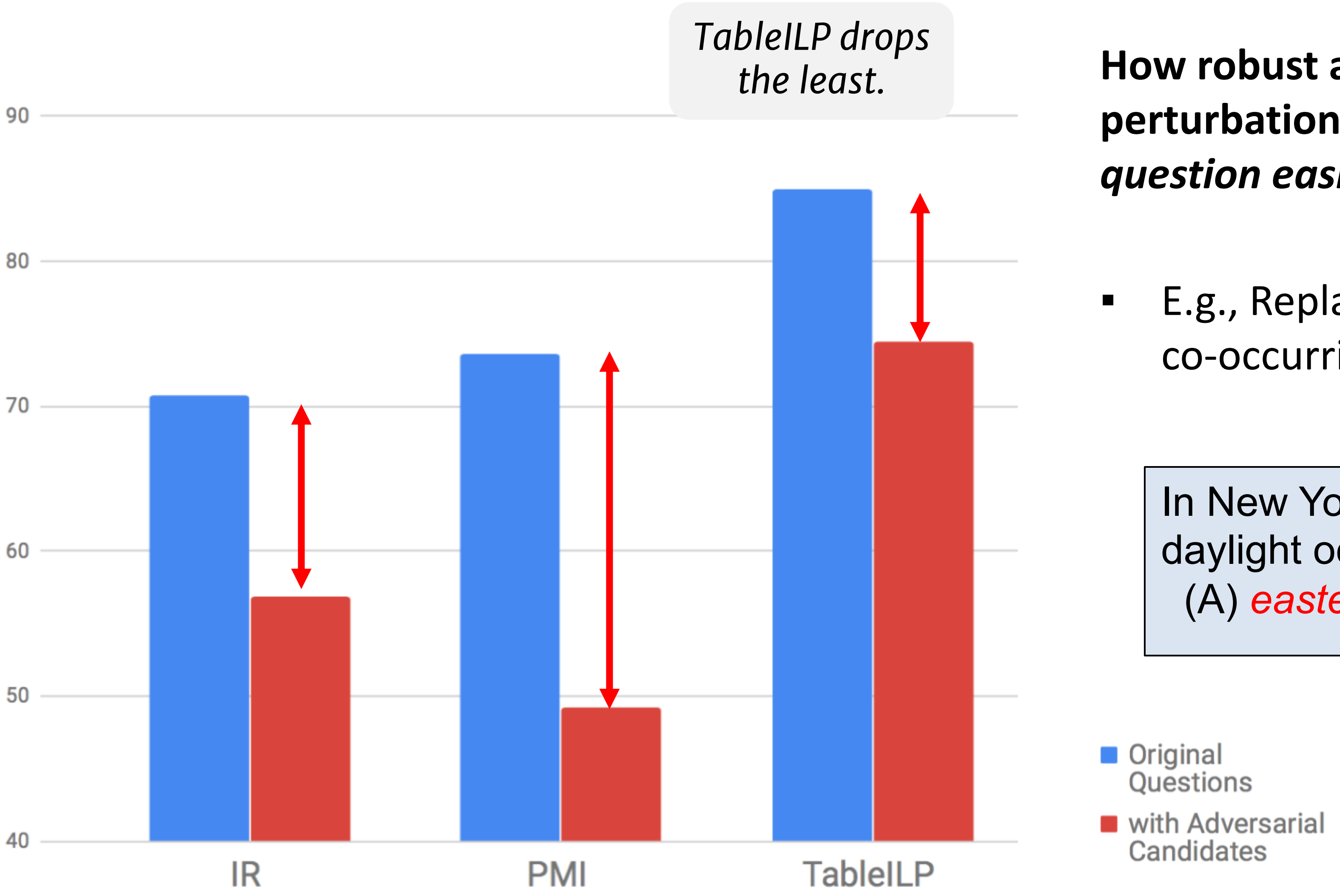


How robust are approaches to simple question perturbations *that would typically make the question easier for a human*?

- 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*

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*

Motivating Example: Circling Back!

- Towards "real understanding" of the phenomenon tested in a question.

In New York State, the longest period of daylight occurs during which month?
In New Zealand, the longest period of daylight occurs during which month?
In New Zealand, the shortest period of daylight occurs during which month?
In New Zealand, the shortest period of night occurs during which month?

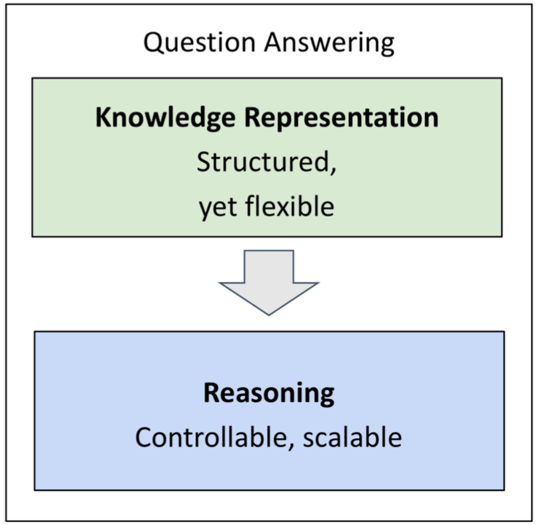
Motivating Example: Circling Back!

- Towards "real understanding" of the phenomenon tested in a question.

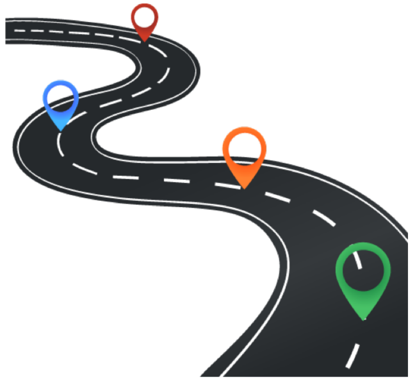
	IR	*PMI*	*TableILP*
In New York State, the longest period of daylight occurs during which month?	*June*	*June*	***June***
In New Zealand, the longest period of daylight occurs during which month?	*March*	*Dec*	***Dec***
In New Zealand, the shortest period of daylight occurs during which month?	*March*	*Dec*	***June***
In New Zealand, the shortest period of night occurs during which month?	*Dec*	*Dec*	***Dec***

Summary, So Far

- Elementary school science tests as a challenge for NLU.
- Knowledge as semi-automatically extracted knowledge.
- Abductive reasoning, to provide the best explanations.
- Showed effective and complementary performances.
- Impacts, since publications:
  - Strong performance on new datasets [Clark et al, Arxiv'2018]
  - Inspired other works [Khot et al, ACL'2017]

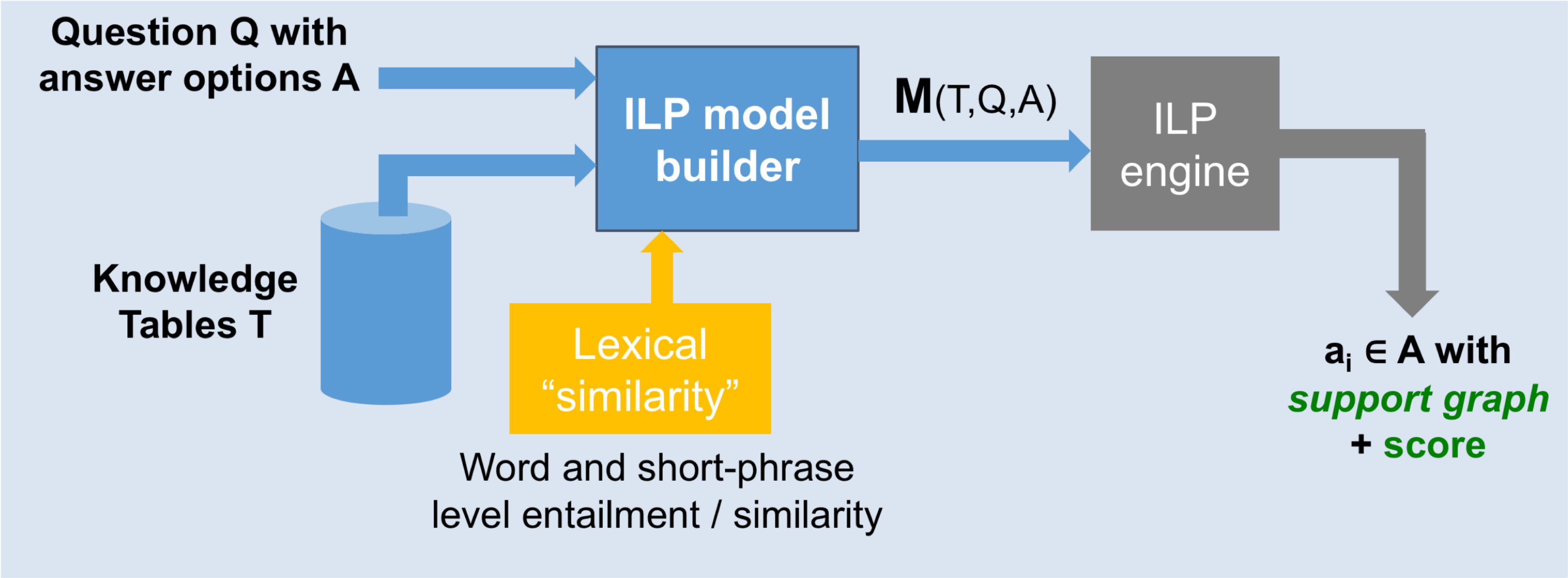


Road Map



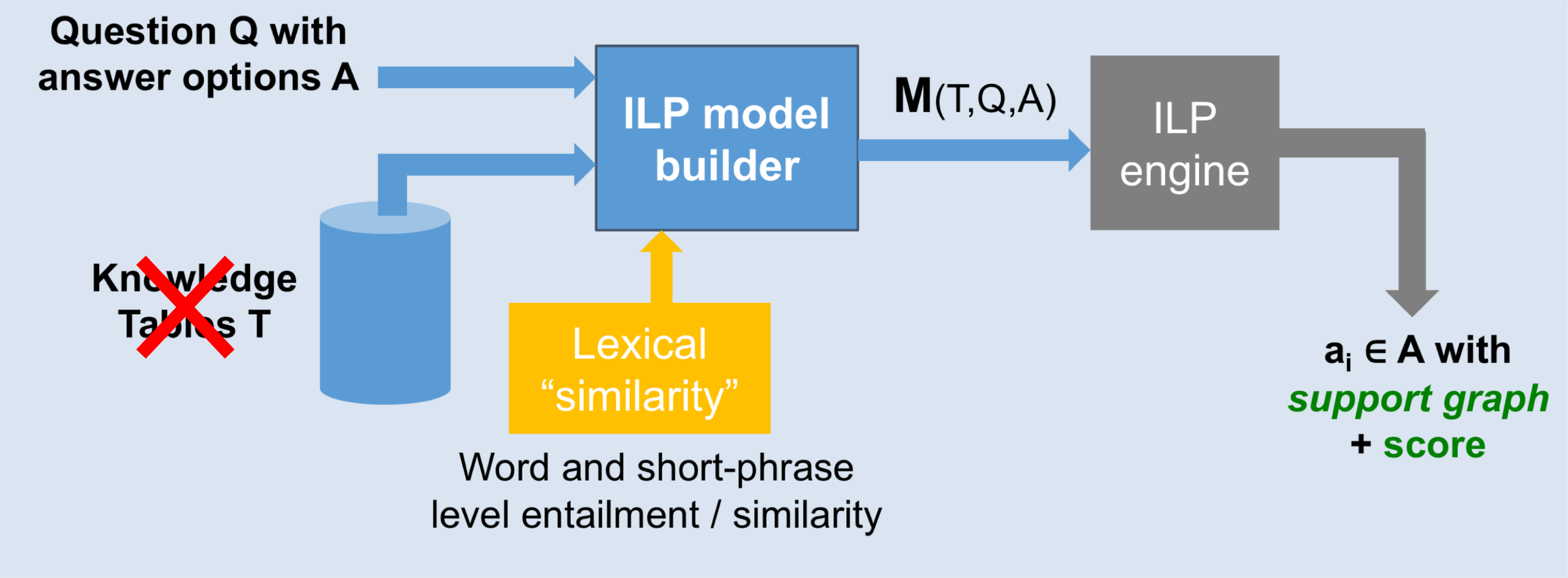
- **Part 1:** Reasoning-Driven System Design
  - QA as Subgraph Optimization on Tabular Knowledge *[IJCAI'16]*
  - QA with Semantic Abstractions of Raw Text *[AAAI'18]*
  - Learning to Pay Attention to Essential Terms in Questions *[CoNLL'17]*

Reasoning With Semantic Abstractions [KKSR'18]



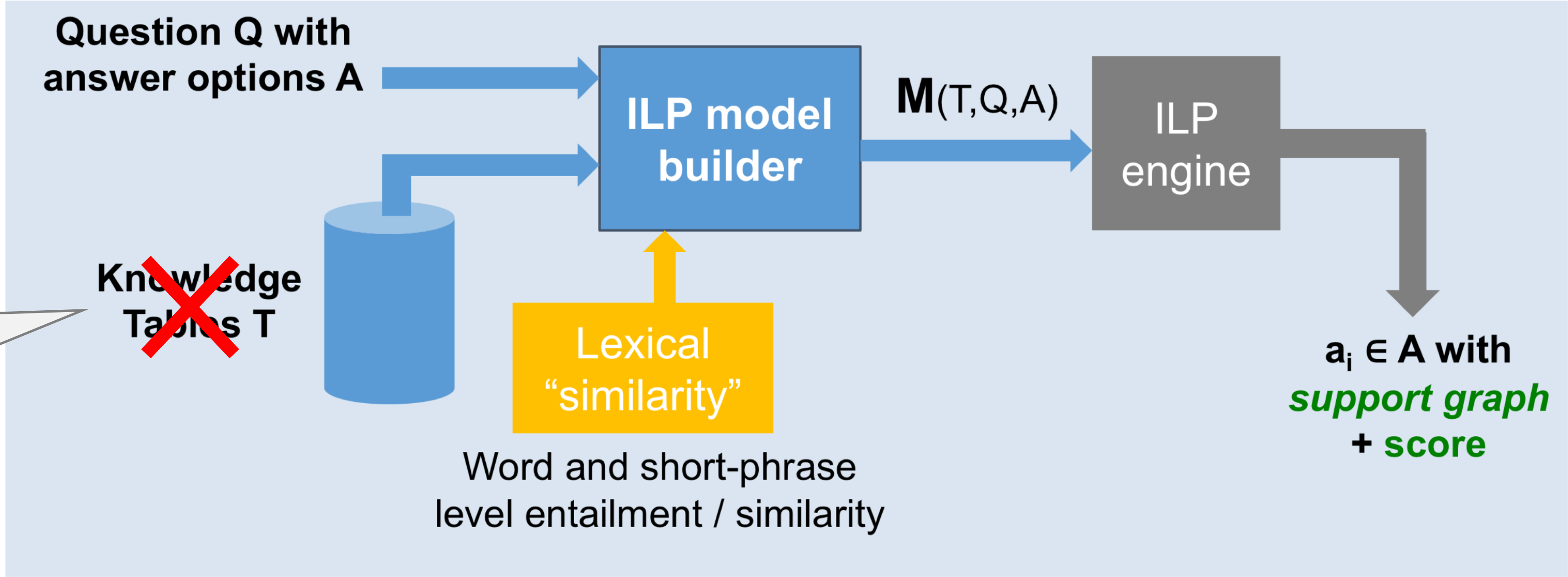
vignettes

Reasoning With Semantic Abstractions [KKSR'18]



vignettes

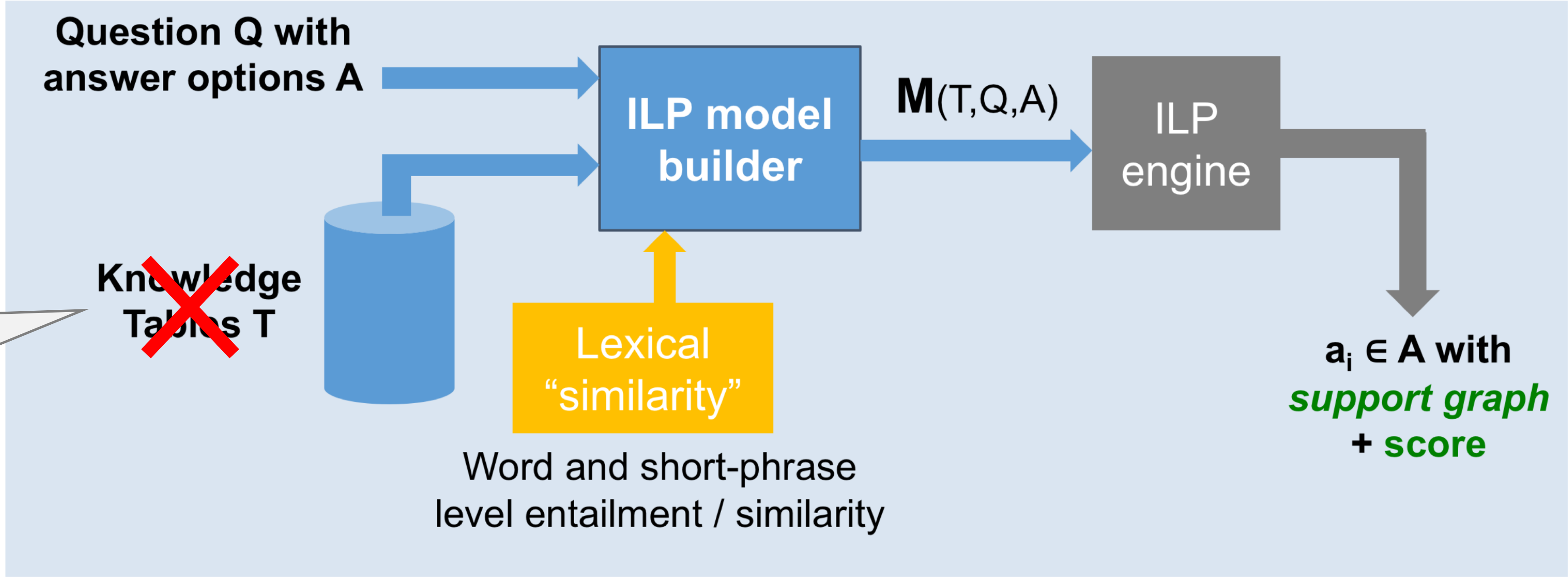
Reasoning With Semantic Abstractions [KKSR'18]



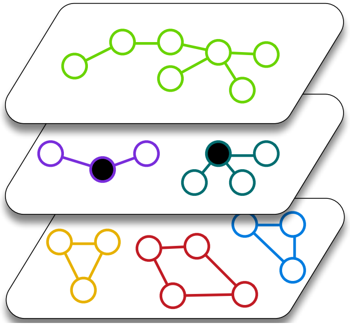
*Knowledge provided as **raw text.***

vignettes

Reasoning With Semantic Abstractions [KKSR'18]



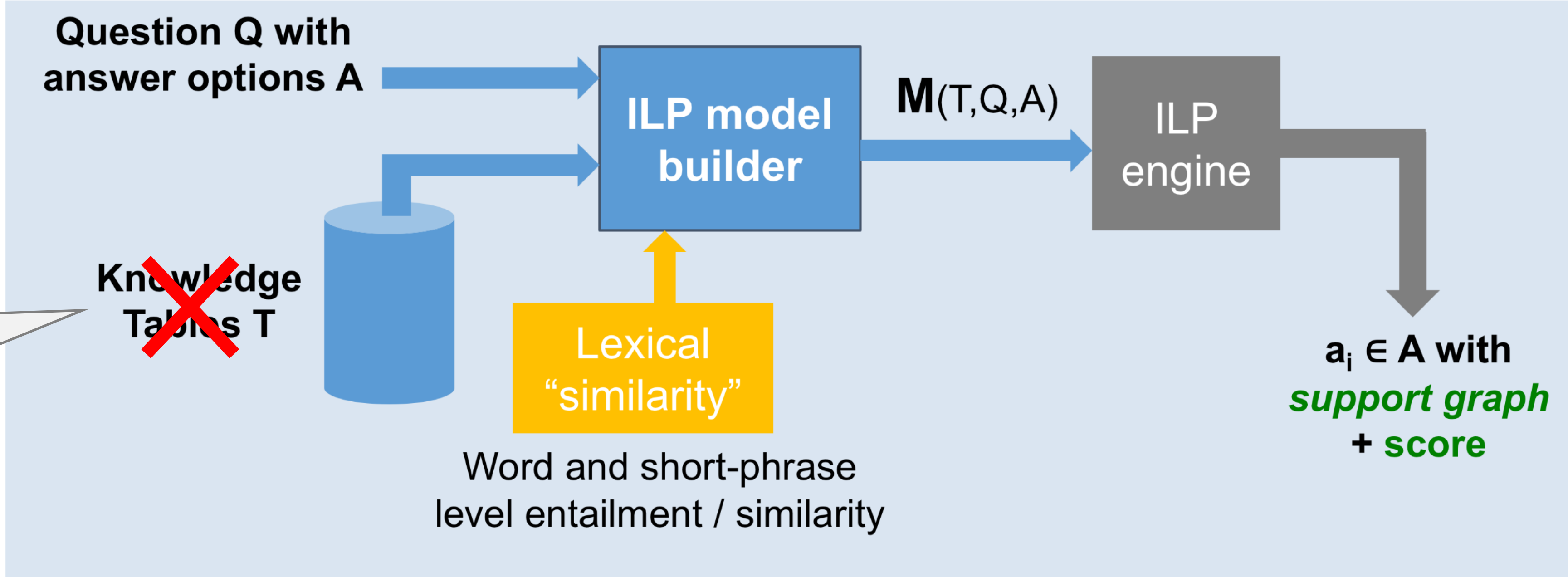
Knowledge provided as ***raw text.***



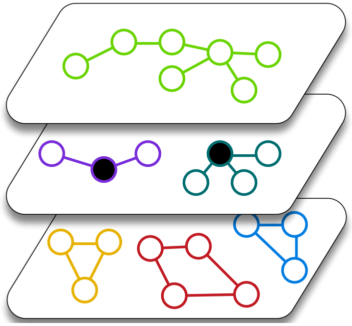
Representing text, as layers of semantic abstractions.

vignettes

Reasoning With Semantic Abstractions [KKSR'18]



Knowledge provided as ***raw text.***

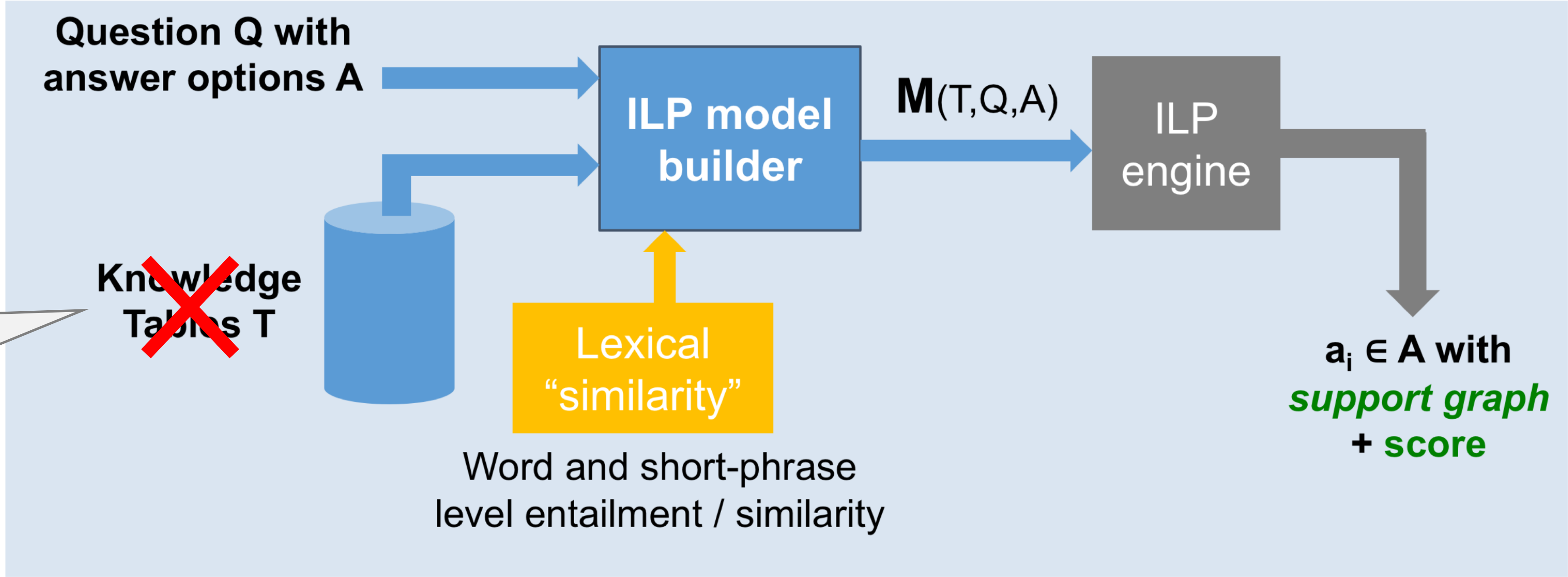


Representing text, as layers of semantic abstractions.

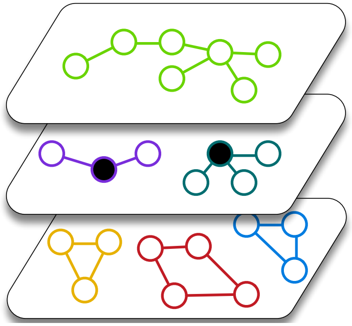
- *Verb-Semantic Roles* [Punyakanok et al, 2008]
- *Preposition-Semantic Roles* [Srikumar & Roth, 2013]
- *Comma-semantic Roles* [Arivazhagan et al, 2016]
- *Coreference* [Chang et al, 2012]

vignettes

Reasoning With Semantic Abstractions [KKSR'18]



Knowledge provided as ***raw text.***



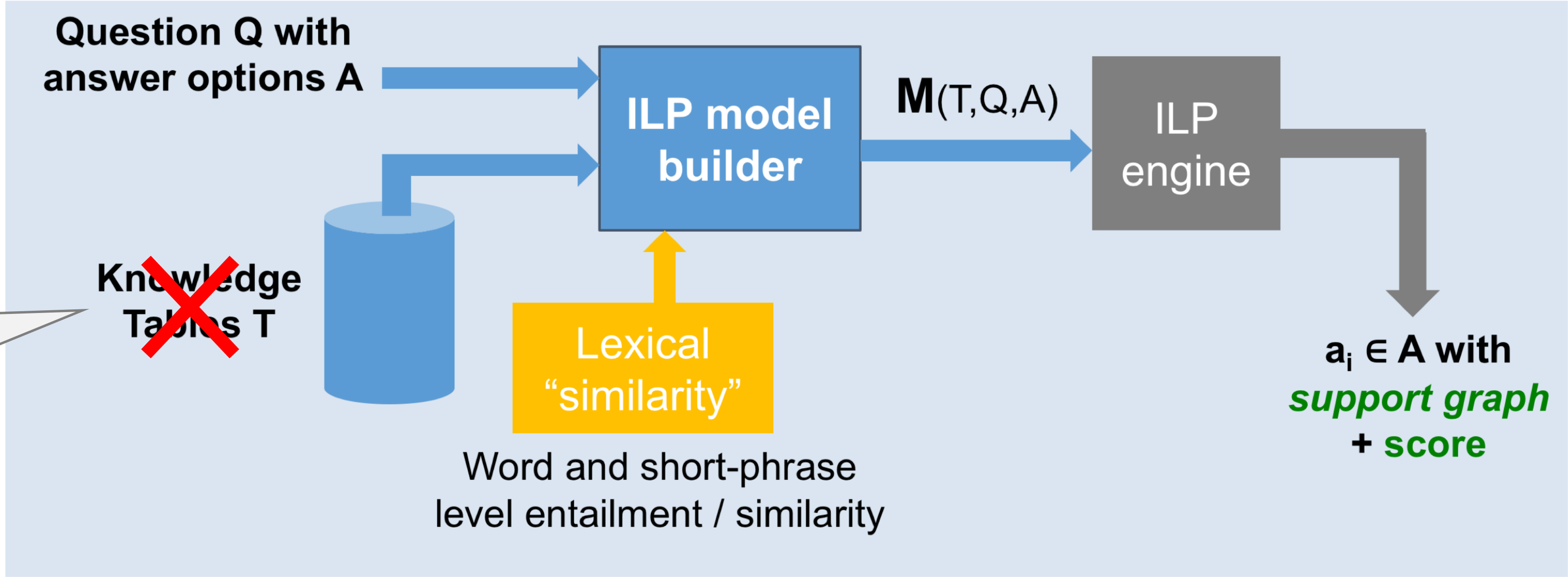
Representing text, as layers of semantic abstractions.

- *Verb-Semantic Roles* [Punyakanok et al, 2008]
- *Preposition-Semantic Roles* [Srikumar & Roth, 2013]
- *Comma-semantic Roles* [Arivazhagan et al, 2016]
- *Coreference* [Chang et al, 2012]

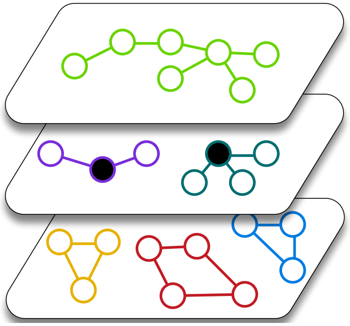
Easier domain transfer!

vignettes

Reasoning With Semantic Abstractions [KKSR'18]



Knowledge provided as ***raw text.***



Representing text, as layers of semantic abstractions.

- *Verb-Semantic Roles* [Punyakanok et al, 2008]
- *Preposition-Semantic Roles* [Srikumar & Roth, 2013]
- *Comma-semantic Roles* [Arivazhagan et al, 2016]
- *Coreference* [Chang et al, 2012]

Easier domain transfer! → Improvements in multiple domains.

vignettes

Learning Essential Terms in Questions [KKSR'17]

Challenge for QA systems: **Is a word in a question *important*, *redundant, or distracting?***

Some animals grows thicker hair as a season changes. This adaptation helps to ________.
(A) find food (B) **keep warmer** (C) grow stronger (D) escape from predators

131

vignettes

Learning Essential Terms in Questions [KKSR'17]

Challenge for QA systems: **Is a word in a question *important, redundant, or distracting?***

Some animals grows thicker hair as a season changes. This adaptation helps to ________.
(A) find food (B) **keep warmer** (C) grow stronger (D) escape from predators

Essentiality in Questions

Learning Essential Terms in Questions [KKSR'17]

Challenge for QA systems: **Is a word in a question *important*, *redundant, or distracting?***

Some animals grows thicker hair as a season changes. This adaptation helps to ________.
(A) find food (B) **keep warmer** (C) grow stronger (D) escape from predators

Essentiality in Questions

2K annotated questions
19K annotated terms

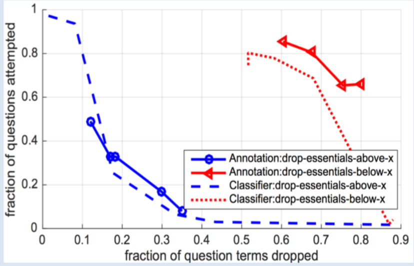
vignettes

Learning Essential Terms in Questions [KKSR'17]

Challenge for QA systems: **Is a word in a question *important, redundant, or distracting?***

Some animals grows thicker hair as a season changes. This adaptation helps to ________.
(A) find food (B) **keep warmer** (C) grow stronger (D) escape from predators

Essentiality in Questions



Important for humans!

amazon mechanicalturk

2K annotated questions
19K annotated terms

vignettes

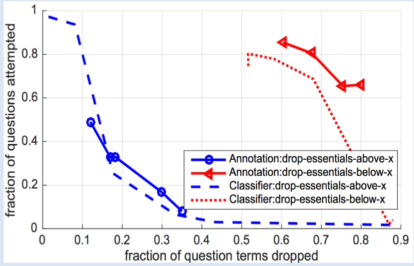
Learning Essential Terms in Questions [KKSR'17]

Challenge for QA systems: **Is a word in a question *important*, *redundant, or distracting?***

Some animals grows thicker hair as a season changes. This adaptation helps to ________.
(A) find food (B) **keep warmer** (C) grow stronger (D) escape from predators

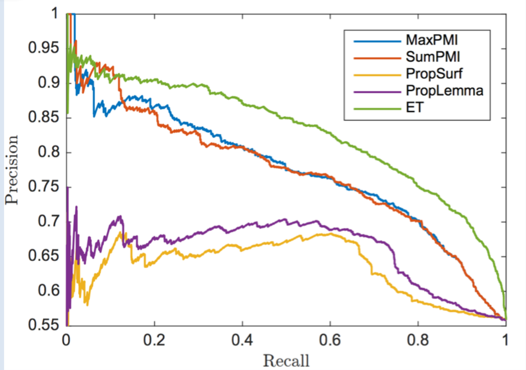
Essentiality in Questions

2K annotated questions
19K annotated terms



Important for humans!

State-of-the-art Essentiality classifier: F1 = 0.8, MAP = 0.9



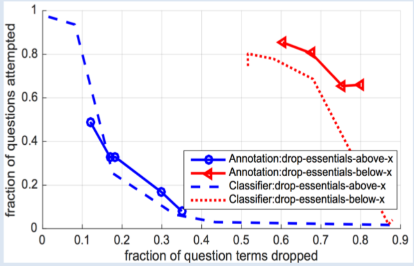
vignettes

Learning Essential Terms in Questions [KKSR'17]

Challenge for QA systems: **Is a word in a question *important*, *redundant, or distracting?***

Some animals grows thicker hair as a season changes. This adaptation helps to ________.
(A) find food (B) **keep warmer** (C) grow stronger (D) escape from predators

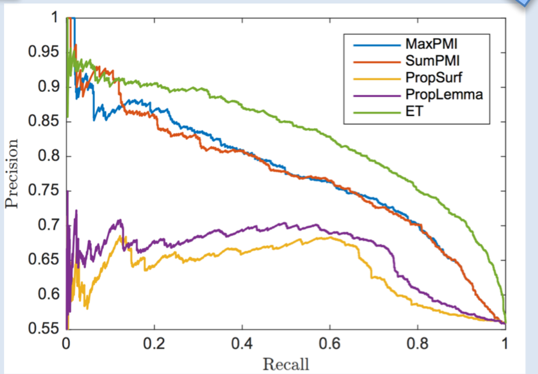
Essentiality in Questions



Important for humans!

2K annotated questions
19K annotated terms

State-of-the-art Essentiality classifier: F1 = 0.8, MAP = 0.9



Dataset	Baseline	With ET
Regents	59.11	**60.85**
AI2Public	57.90	**59.10**
RegtsPertd	61.84	**66.84**

Up to 5% increase in end-to-end QA performance

Road Map

- **Part 2:** Moving the Peaks Higher: More Challenging Datasets
    - A QA Benchmark for Temporal Common Sense *[Submitted]*
    - A QA Benchmark for Reasoning on Multiple Sentences *[NAACL'18]*

The Recent State of the Field: Good News

- Many large QA datasets [Rajpurkar at al, 2016; others]
- Successes, with neural nets
  - Faster computers (e.g., GPUs)
  - New computational modules (e.g., Attentions)
  - More data

The Recent State of the Field: Good News

- Many large QA datasets [Rajpurkar at al, 2016; others]
- Successes, with neural nets
  - Faster computers (e.g., GPUs)
  - New computational modules (e.g., Attentions)
  - More data

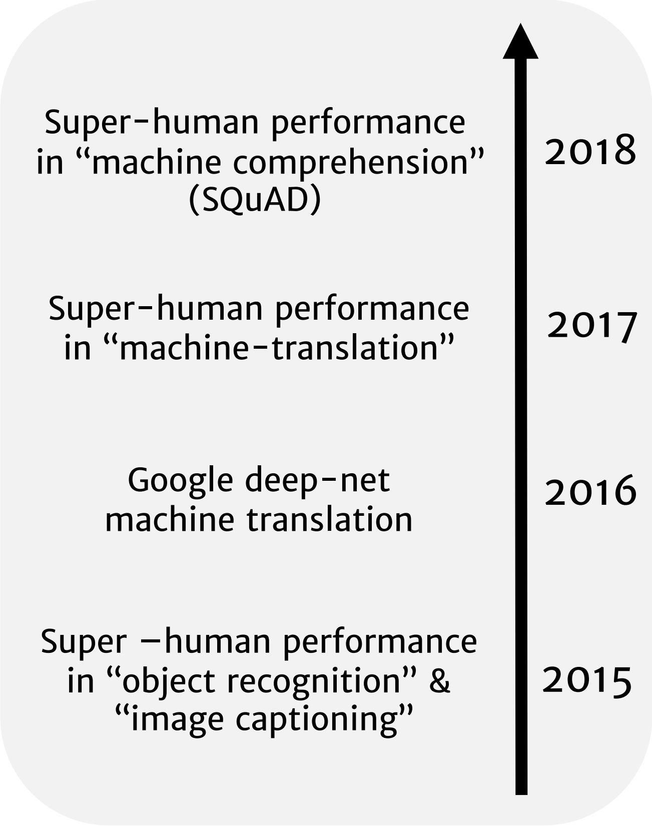
The Recent State of the Field: Good News

- Many large QA datasets [Rajpurkar at al, 2016; others]
- Successes, with neural nets
  - Faster computers (e.g., GPUs)
  - New computational modules (e.g., Attentions)
  - More data

Super-human performance in “machine comprehension” (SQuAD)	2018
Super-human performance in “machine-translation”	2017
Google deep-net machine translation	2016
Super –human performance in “object recognition” & “image captioning”	2015

The Recent State of the Field: Good News

- Many large QA datasets [Rajpurkar at al, 2016; others]
- Successes, with neural nets
  - Faster computers (e.g., GPUs)
  - New computational modules (e.g., Attentions)
  - More data



The Recent State of the Field: Good News

- Many large QA datasets [Rajpurkar at al, 2016; others]
- Successes, with neural nets
  - Faster computers (e.g., GPUs)
  - New computational modules (e.g., Attentions)
  - More data

Super-human performance in “machine comprehension” (SQuAD) — 2018

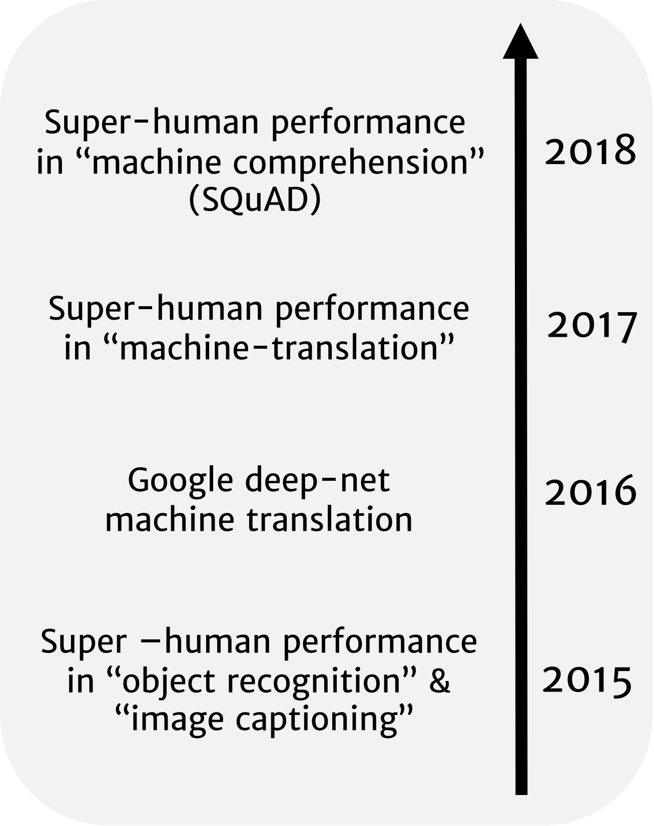
Super-human performance in “machine-translation” — 2017

Google deep-net machine translation — 2016

Super –human performance in “object recognition” & “image captioning” — 2015

The Recent State of the Field: Good News

- Many large QA datasets [Rajpurkar at al, 2016; others]
- Successes, with neural nets
  - Faster computers (e.g., GPUs)
  - New computational modules (e.g., Attentions)
  - More data



The Recent State of the Field: Bad News

The Recent State of the Field: Bad News

- The systems easily break-down. [K at al, 2016; Jia et al, 2017; Belinkov et al, 2018; others]
- Many problems with no significant success: Math word problems; dialogue; many others.
  - Some are not even defined yet.
- Discoveries are more about tasks (datasets).
- The urge to scale up datasets has biased in certain angles and limited their diversity.

The Recent State of the Field: Bad News

- The systems easily break-down. [K at al, 2016; Jia et al, 2017; Belinkov et al, 2018; others]
- Many problems with no significant success: Math word problems; dialogue; many others.
    - Some are not even defined yet.
- Discoveries are more about tasks (datasets).
- The urge to scale up datasets has biased in certain angles and limited their diversity.

The Recent State of the Field: Bad News

- The systems easily break-down. [K at al, 2016; Jia et al, 2017; Belinkov et al, 2018; others]
- Many problems with no significant success: Math word problems; dialogue; many others.
    - Some are not even defined yet.
- Discoveries are more about tasks (datasets).
- The urge to scale up datasets has biased in certain angles and limited their diversity.

The Recent State of the Field: Bad News

- The systems easily break-down. [K at al, 2016; Jia et al, 2017; Belinkov et al, 2018; others]
- Many problems with no significant success: Math word problems; dialogue; many others.
  - Some are not even defined yet.
- Discoveries are more about tasks (datasets). [Darwiche, 2017]
- The urge to scale up datasets has biased in certain angles and limited their diversity.

The Recent State of the Field: Bad News

- The systems easily break-down. [K at al, 2016; Jia et al, 2017; Belinkov et al, 2018; others]
- Many problems with no significant success: Math word problems; dialogue; many others.
  - Some are not even defined yet.
- Discoveries are more about tasks (datasets). [Darwiche, 2017]
- The urge to scale up datasets has biased in certain angles and limited their diversity.

Moving the Peaks Higher

Moving the Peaks Higher

Goal here:

Define and create challenges not addressed by the community.

Challenges that require external knowledge, common-sense, complex reasoning, etc.

Moving the Peaks Higher

Goal here:

Define and create challenges not addressed by the community.

Challenges that require external knowledge, common-sense, complex reasoning, etc.

Including our systems

Moving the Peaks Higher

Goal here:

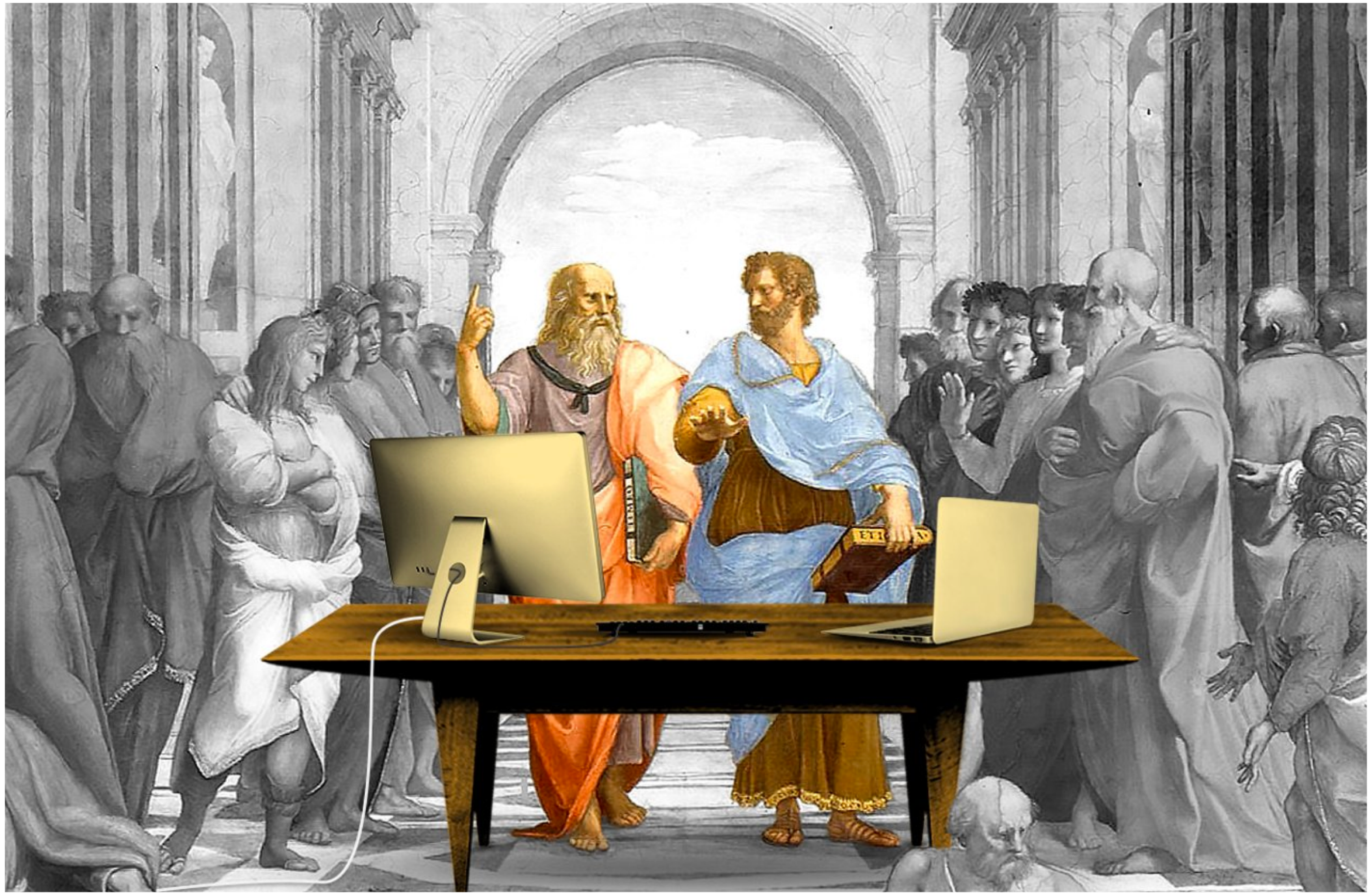
Define and create challenges not addressed by the community.

Challenges that require external knowledge, common-sense, complex reasoning, etc.

Including our systems

Road Map

- **Part 2:** Moving the Peaks Higher: More Challenging Datasets
  - A QA Benchmark for Temporal Common Sense *[Submitted]*
  - A QA Benchmark for Reasoning on Multiple Sentences *[NAACL'18]*



Did Aristotle have a laptop?

[Valiant, ?]

Google

Did Aristotle have a laptop?

All News Shopping Videos Images More Settings Tools

About 1,100,000 results (0.41 seconds)

Aristotle's Laptop | Series on Machine Consciousness - World Scientific
https://www.worldscientific.com/worldscibooks/10.1142/8113
This book is about a scientific ingredient that **was** not available to **Aristotle**: the science of information. Would the course of the philosophy of the mind **have** been ...

Aristotle's Laptop: The discovery of out informational mind | Request PDF
https://www.researchgate.net/.../259464884_Aristotle's_Laptop_The_discovery_of_out_...
Aristotle's convincing philosophy is likely to **have** shaped (even indirectly) many ... mind **have** been different **had Aristotle** pronounced that the matter of mind **was** ...

Aristotle's Laptop eBook by Igor Aleksander - 9789814425629 ...
https://www.kobo.com › ... › Advanced Computing › Artificial Intelligence
Read "**Aristotle's Laptop** The Discovery of our Informational Mind" by Igor ... **Aristotle's** convincing philosophy is likely to **have** shaped (even indirectly) ... **have** been different **had Aristotle** pronounced that the matter of mind **was** information?

Types of “Knowledge”

- [Loosely-defined] types of knowledge
  - Encyclopedic knowledge
  - Commonsense knowledge

Types of "Knowledge"

- [Loosely-defined] types of knowledge
  - Encyclopedic knowledge
    - *The capital city of Venezuela is Caracas.*
    - *Obama was born in Honolulu, HI.*
  - Commonsense knowledge

Types of “Knowledge”

- [Loosely-defined] types of knowledge
  - Encyclopedic knowledge
    - *The capital city of Venezuela is Caracas.*
    - *Obama was born in Honolulu, HI.*
  - Commonsense knowledge
    - *If you lived thousands of years ago, you’re unlikely to be alive now.*
    - *Laptops didn’t exist, before they were invented.*

Types of "Knowledge"

- [Loosely-defined] types of knowledge
  - Encyclopedic knowledge
    - *The capital city of Venezuela is Caracas.*
    - *Obama was born in Honolulu, HI.*
  - Commonsense knowledge
    - *If you lived thousands of years ago, you're unlikely to be alive now.*
    - *Laptops didn't exist, before they were invented.*

Focus of this work

Common Sense: A Short History

Common Sense: A Short History

- Many early works
  - Since the early days of AI
  - Ambitious projects
- Recent years:
  - Winograd Schema Challenge

Common Sense: A Short History

- Many early works
  - Since the early days of AI [McCarthy, 68; Charniak, 77; others]
  - Ambitious projects
- Recent years:
  - Winograd Schema Challenge

Common Sense: A Short History

- Many early works
  - Since the early days of AI [McCarthy, 68; Charniak, 77; others]
  - Ambitious projects [Lenat, 85; others]
- Recent years:
  - Winograd Schema Challenge

Common Sense: A Short History

- Many early works
  - Since the early days of AI [McCarthy, 68; Charniak, 77; others]
  - Ambitious projects [Lenat, 85; others]
- Recent years:
  - Winograd Schema Challenge [Levesque, 2014; Peng, K, Roth, 2015]

Common Sense: A Short History

- Many early works
  - Since the early days of AI [McCarthy, 68; Charniak, 77; others]
  - Ambitious projects [Lenat, 85; others]
- Recent years:
  - Winograd Schema Challenge [Levesque, 2014; Peng, K, Roth, 2015]

"Jack pulled up a picture of Aristotle on his laptop"

"Jack pulled up a picture of Aristotle from his biography"

Common Sense: A Short History

- Many early works
  - Since the early days of AI [McCarthy, 68; Charniak, 77; others]
  - Ambitious projects [Lenat, 85; others]
- Recent years:
  - Winograd Schema Challenge [Levesque, 2014; Peng, K, Roth, 2015]

"Jack pulled up a picture of Aristotle on his laptop"

"Jack pulled up a picture of Aristotle from his biography"

Incentivizing commonsense understanding as a high-level and well-defined task.

Temporal Common Sense

- Understanding “time” is a key ability in many NLU tasks.

Temporal Common Sense

- Understanding “time” is a key ability in many NLU tasks.

“Going to barbershop” takes a couple of **hours**

“Going to college” takes a couple of **years**

Event duration

Temporal Common Sense

- Understanding "time" is a key ability in many NLU tasks.

"Going to barbershop" takes a couple of **hours**

"Going to college" takes a couple of **years**

Event duration

"Take a trip to Africa" happens once in a few **years**

"Take a trip to parent's" happens every few **weeks** or **months**

Event frequency

Temporal Common Sense

- Understanding "time" is a key ability in many NLU tasks.

Examples	Type
"Going to barbershop" takes a couple of **hours** "Going to college" takes a couple of **years**	Event duration
"Take a trip to Africa" happens once in a few **years** "Take a trip to parent's" happens every few **weeks** or **months**	Event frequency
"Going to work" usually happens around **morning** time "Going to a bar" usually happens around **evening/night** time	Typical time

Temporal Common Sense

- Understanding "time" is a key ability in many NLU tasks.

"Going to barbershop" takes a couple of **hours**

"Going to college" takes a couple of **years**

Event duration

"Take a trip to Africa" happens once in a few **years**

"Take a trip to parent's" happens every few **weeks** or **months**

Event frequency

"Going to work" usually happens around **morning** time

"Going to a bar" usually happens around **evening/night** time

Typical time

Goal of this section:

QA dataset that requires temporal commonsense.

Temporal Common Sense as QA

- A dataset of natural language questions
  - Questions about a “temporal” understanding
    - About 1k questions and 8k candidate answers
    - 5 temporal phenomena:
      - Event duration, event frequency, absolute time point, event ordering, stationary vs transient

Scenario: *I clapped her shoulder to show I was not laughing at her.*
Question: *How long did they laugh?*
Candidates:
(A) for a few days (B) 20 minutes **(C) for a few minutes**

Temporal Common Sense as QA

- A dataset of natural language questions
  - Questions about a “temporal” understanding
    - About 1k questions and 8k candidate answers
    - 5 temporal phenomena:
      - Event duration, event frequency, absolute time point, event ordering, stationary vs transient

Scenario: *With the amount of money Diggler was making he was able to support both his and Rothchild's addictions.*
Question: *What time did Diggler go to work?*
Candidates:
(A) at eight in the late night (B) he leaves around 3 am **(C) he leaves around 8 am**

Temporal Common Sense as QA

- A dataset of natural language questions
  - Questions about a "temporal" understanding
    - About 1k questions and 8k candidate answers
    - 5 temporal phenomena:
      - Event duration, event frequency, absolute time point, event ordering, stationary vs transient

Scenario: *She checked the kitchen, but didn't find anything missing there except for a clock.*
Question: *Why was she checking for missing items?*
Candidates:
(A) her window had been broken (B) someone tied her hands

Temporal Common Sense as QA

- A dataset of natural language questions
  - Questions about a "temporal" understanding
    - About 1k questions and 8k candidate answers
    - 5 temporal phenomena:
      - Event duration, event frequency, absolute time point, event ordering, stationary vs transient

Scenario: *They then took a boat to Africa and Asia, where they went on a trip through the mountains.*
Question: *How often do they go on trips?*
Candidates:

(A) every night **(B) once a year**
(C) twice a year (D) once a week

TacoQA: Construction Overview

Find the qualified annotators

Define the task and the desired questions

TacoQA: Construction Overview

amazon mechanicalturk

Find the qualified annotators

Define the task and the desired questions

Generate questions, given scenarios, for each temporal phenomena

TacoQA: Construction Overview

Find the qualified annotators

Define the task and the desired questions

Generate questions, given scenarios, for each temporal phenomena

Q: How often do they go on trips?
A: Once a year

TacoQA: Construction Overview

Find the qualified annotators

Verify the question quality and their temporal category

Define the task and the desired questions

Generate questions, given scenarios, for each temporal phenomena

Q: How often do they go on trips?
A: Once a year

TacoQA: Construction Overview

Define the task and the desired questions

amazon mechanicalturk

Find the qualified annotators

Generate questions, given scenarios, for each temporal phenomena

Q: How often do they go on trips?
A: Once a year

Verify the question quality and their temporal category

Candidate answer expansion

Once a month
Once a decade
Twice a year
Never
Rarely

TacoQA: Construction Overview

Define the task and the desired questions

amazon mechanicalturk

Find the qualified annotators

Generate questions, given scenarios, for each temporal phenomena

Q: How often do they go on trips?
A: Once a year

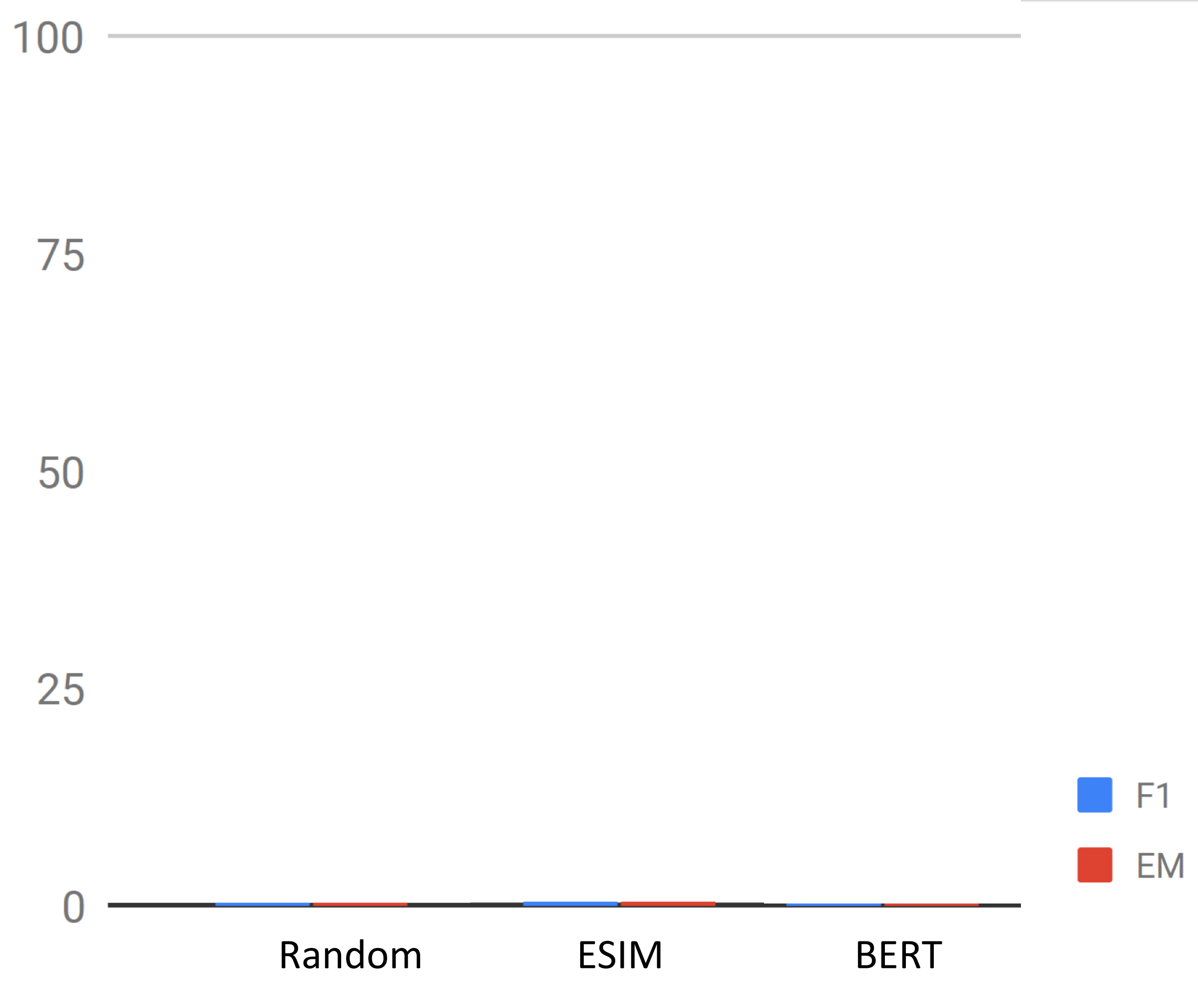
Verify the question quality and their temporal category

Candidate answer expansion

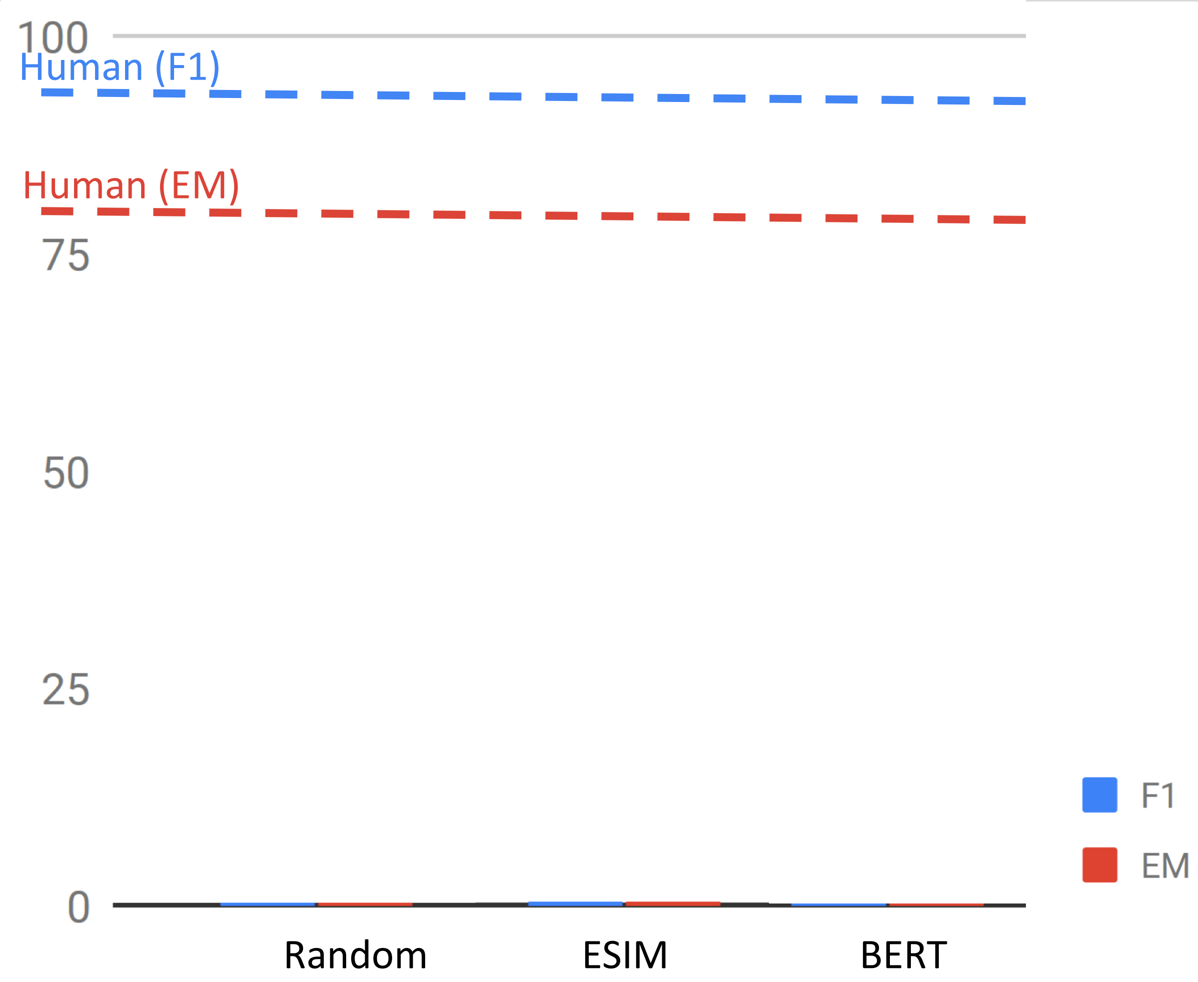
Once a month
Once a decade
Twice a year
Never
Rarely

Answer verification

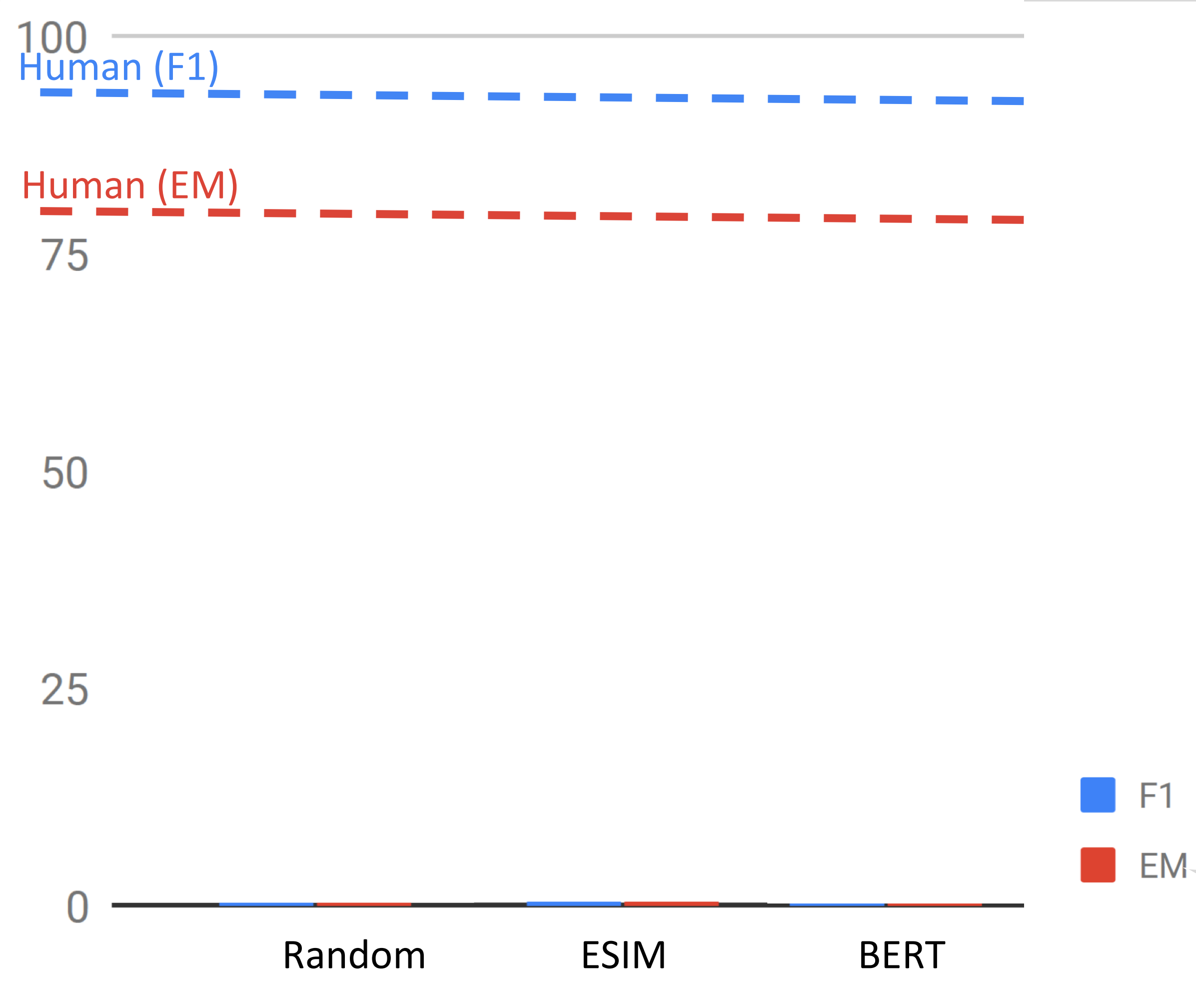
Experimental Results [ZKNR, under review]



Experimental Results [ZKNR, under review]

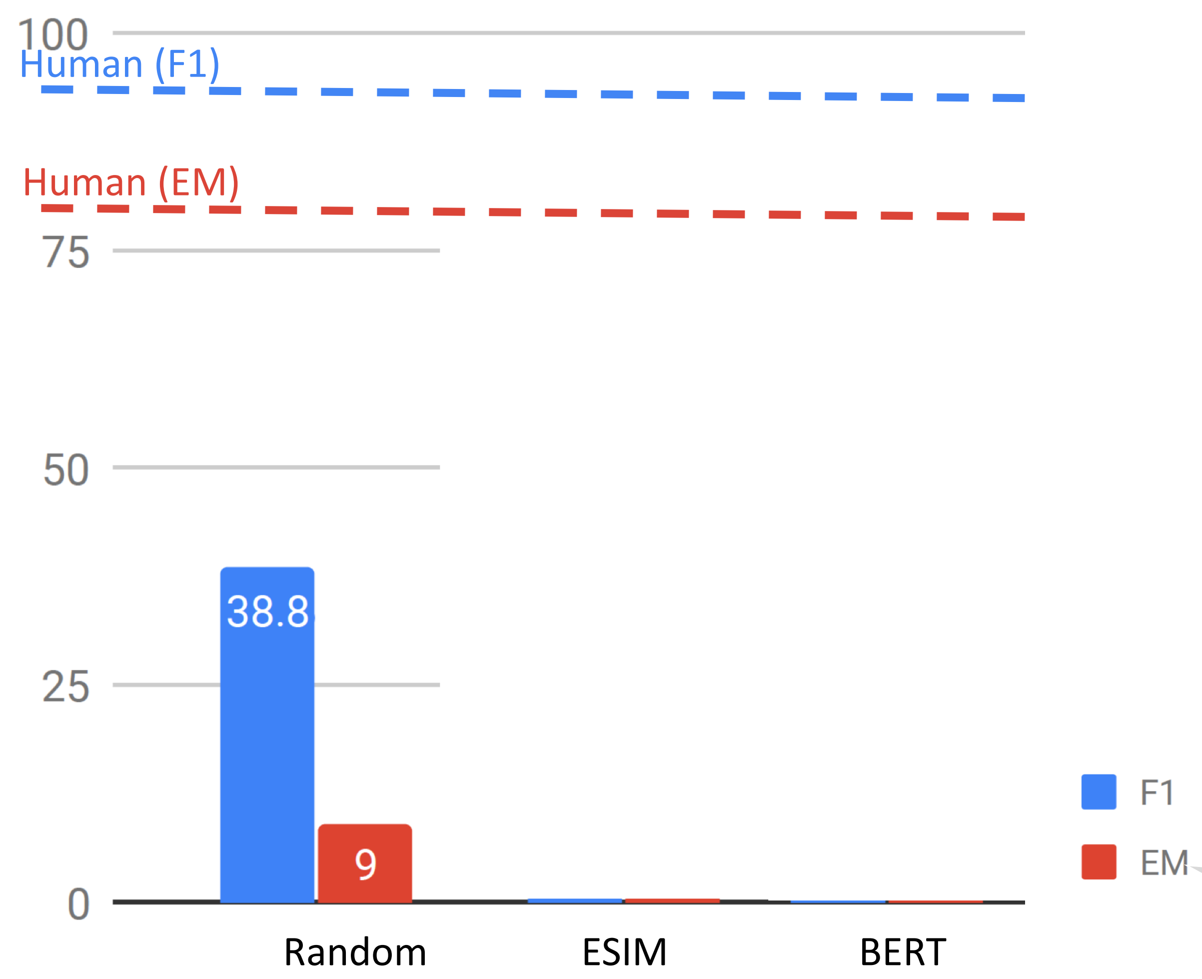


Experimental Results [ZKNR, under review]



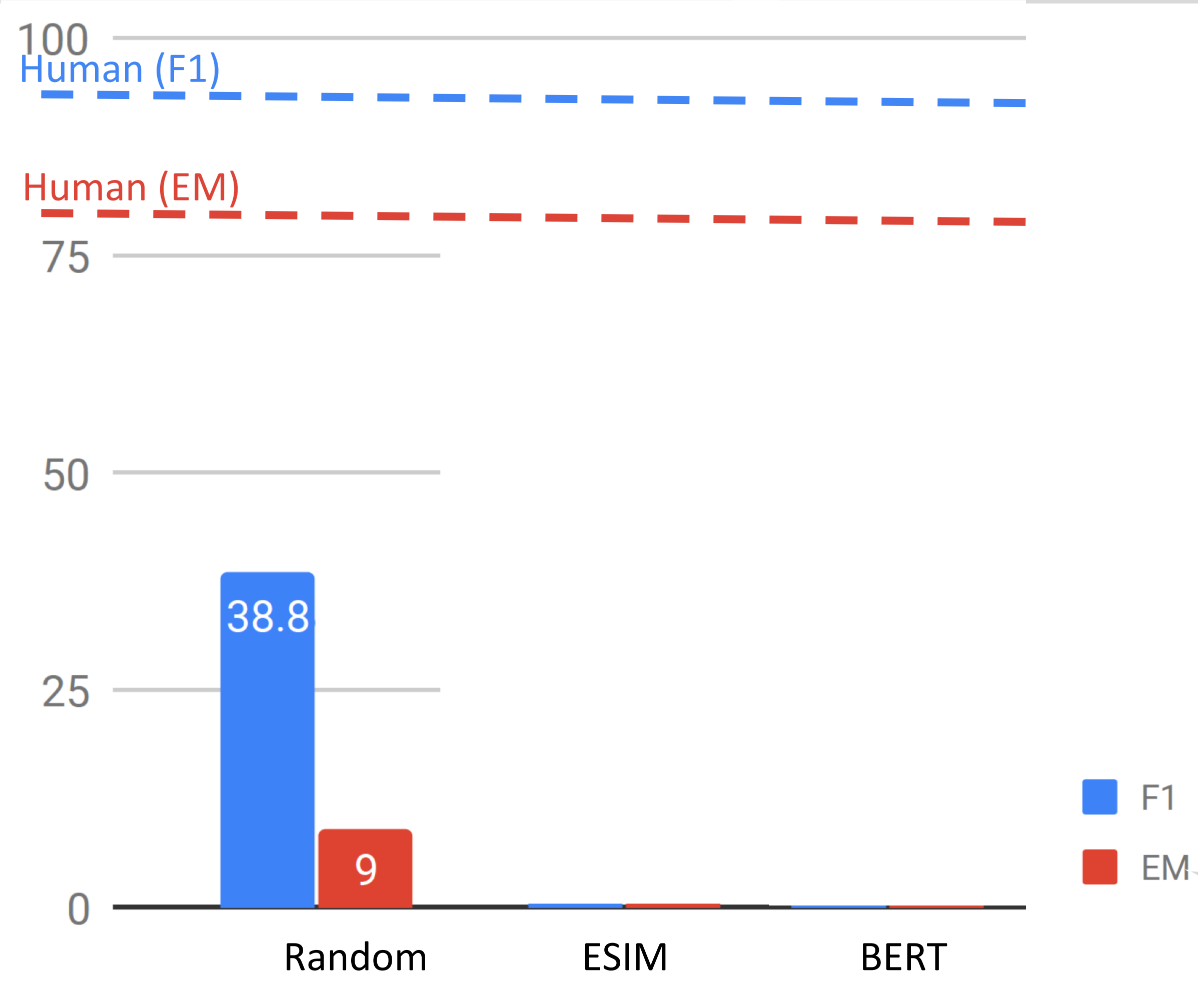
A systems gets credit only if, it gets **all** the candidates right.

Experimental Results [ZKNR, under review]



A systems gets credit only if, it gets **all** the candidates right.

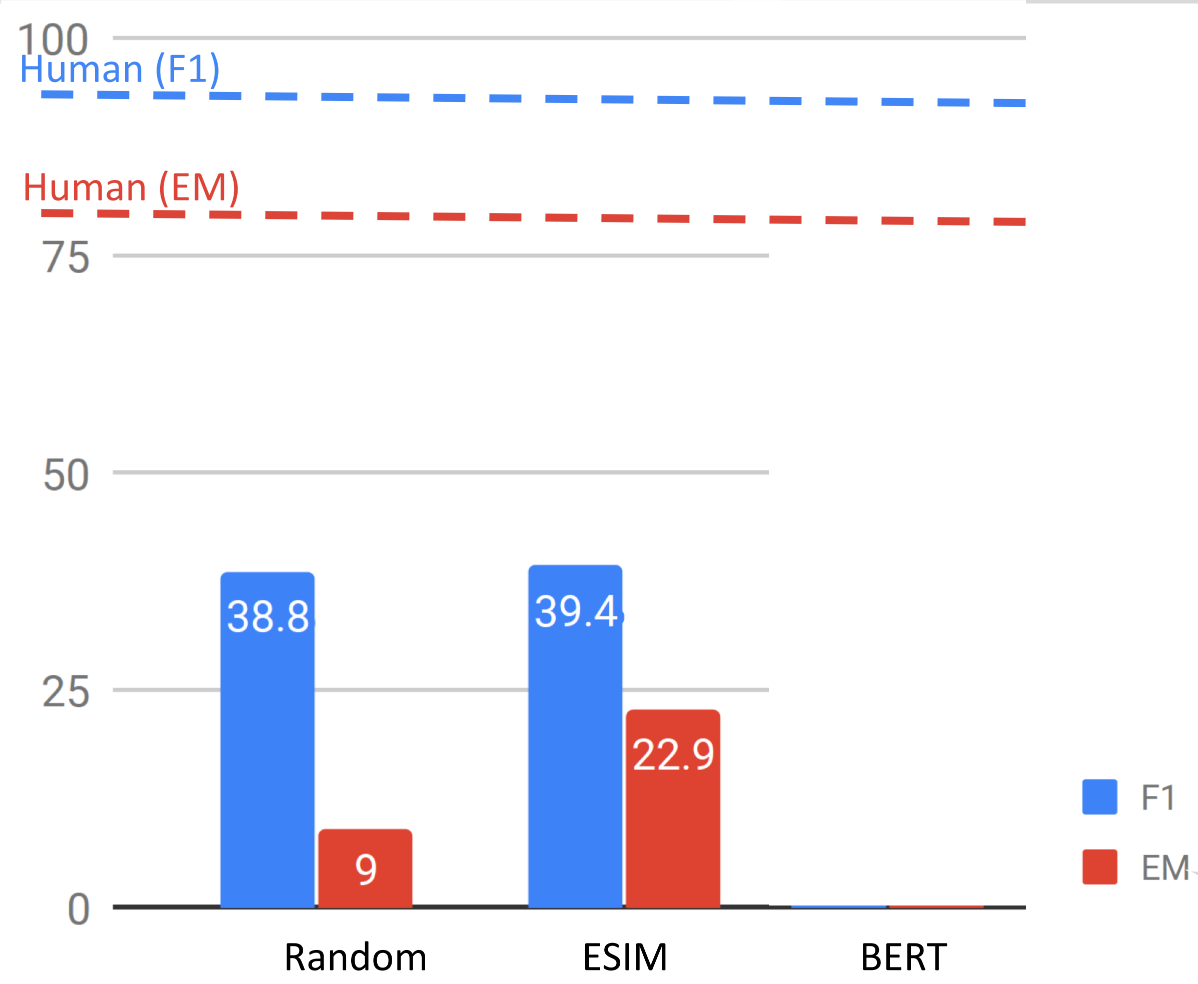
Experimental Results [ZKNR, under review]



A systems gets credit only if, it gets **all** the candidates right.

Supervised-learning (LSTM)
[Chen et al, 2017]

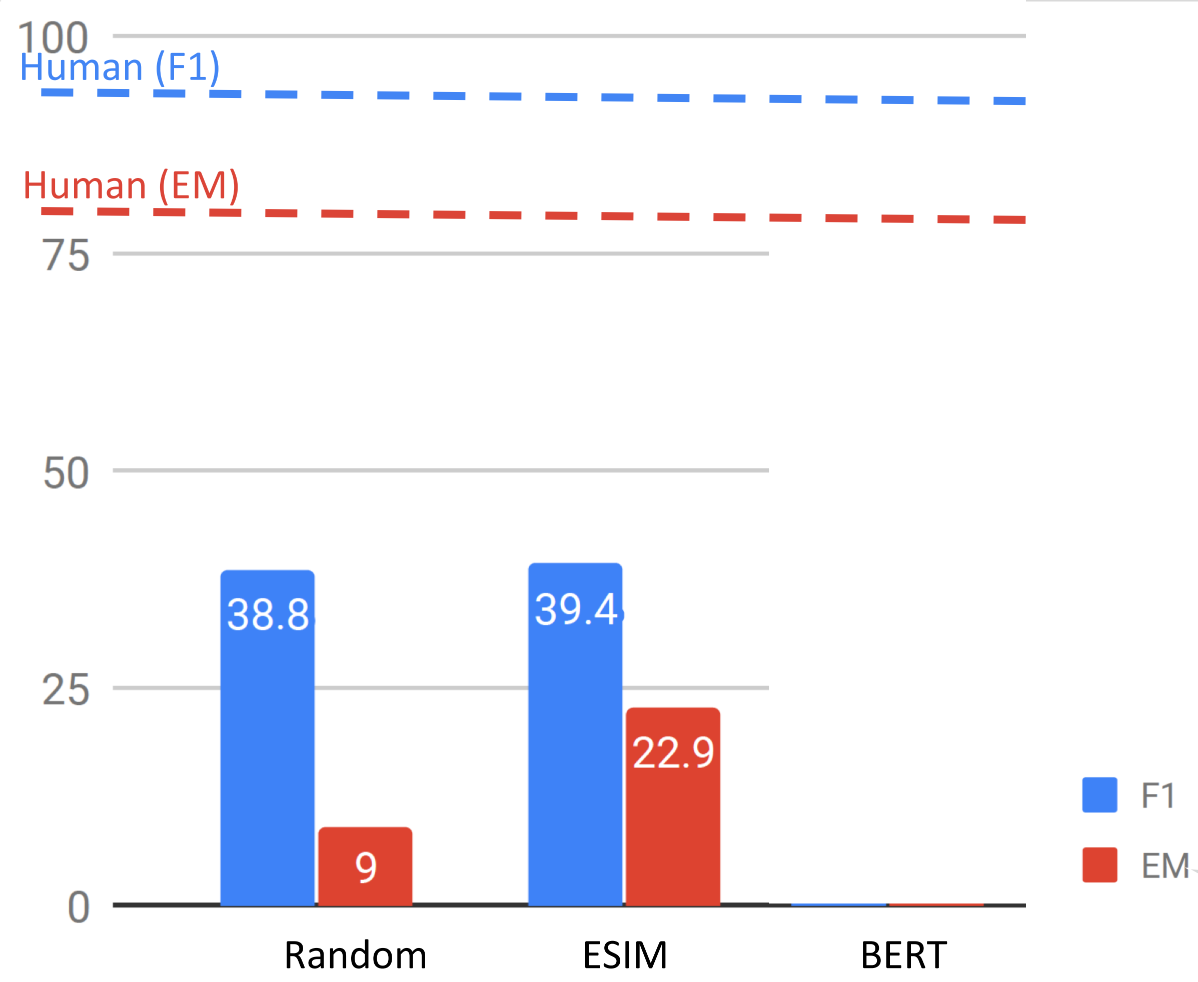
Experimental Results [ZKNR, under review]



A systems gets credit only if, it gets **all** the candidates right.

Supervised-learning (LSTM)
[Chen et al, 2017]

Experimental Results [ZKNR, under review]

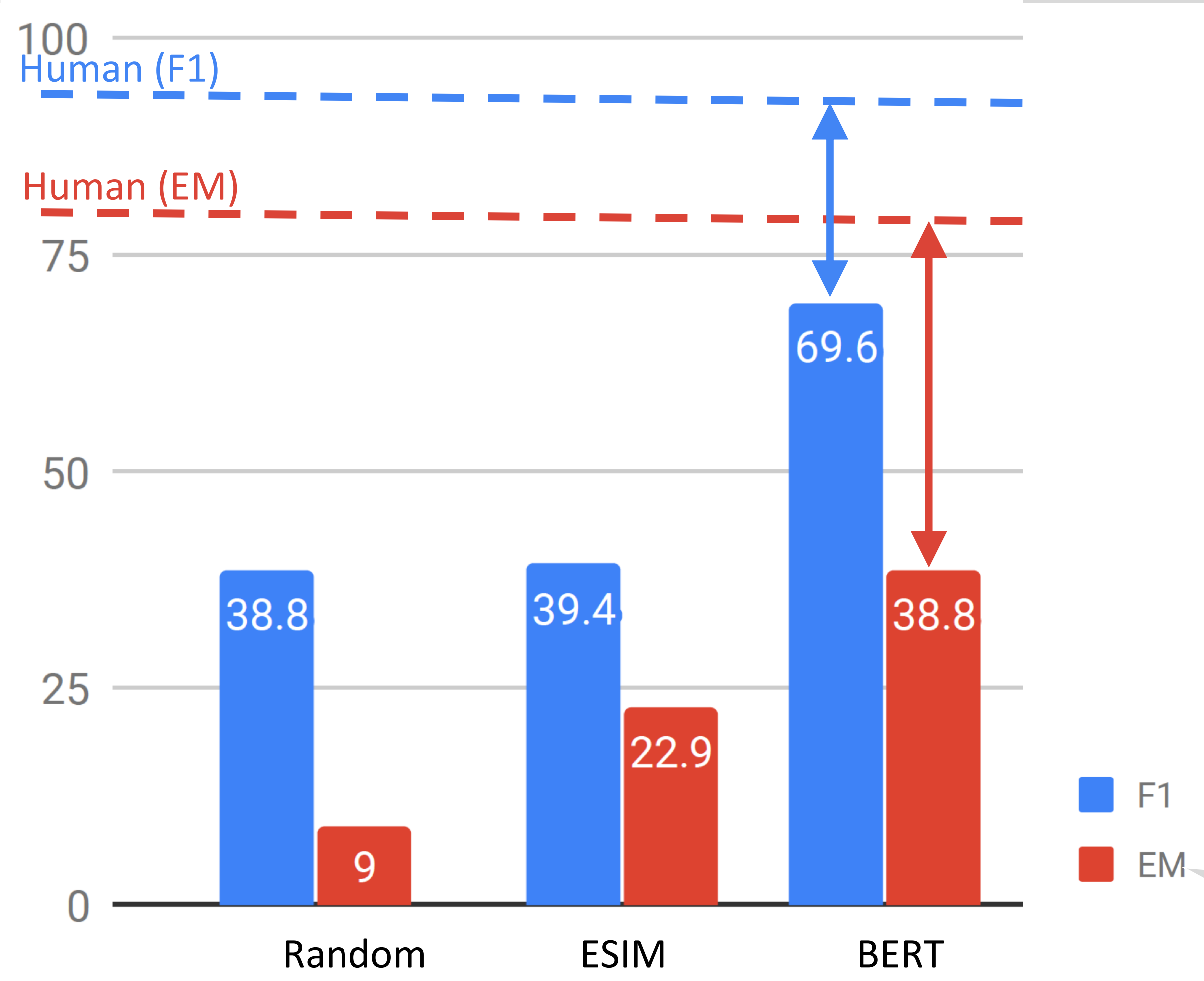


A systems gets credit only if, it gets **all** the candidates right.

Supervised-learning (LSTM)
[Chen et al, 2017]

Pre-trained contextualized model
[Devlin et al, 2018]

Experimental Results [ZKNR, under review]

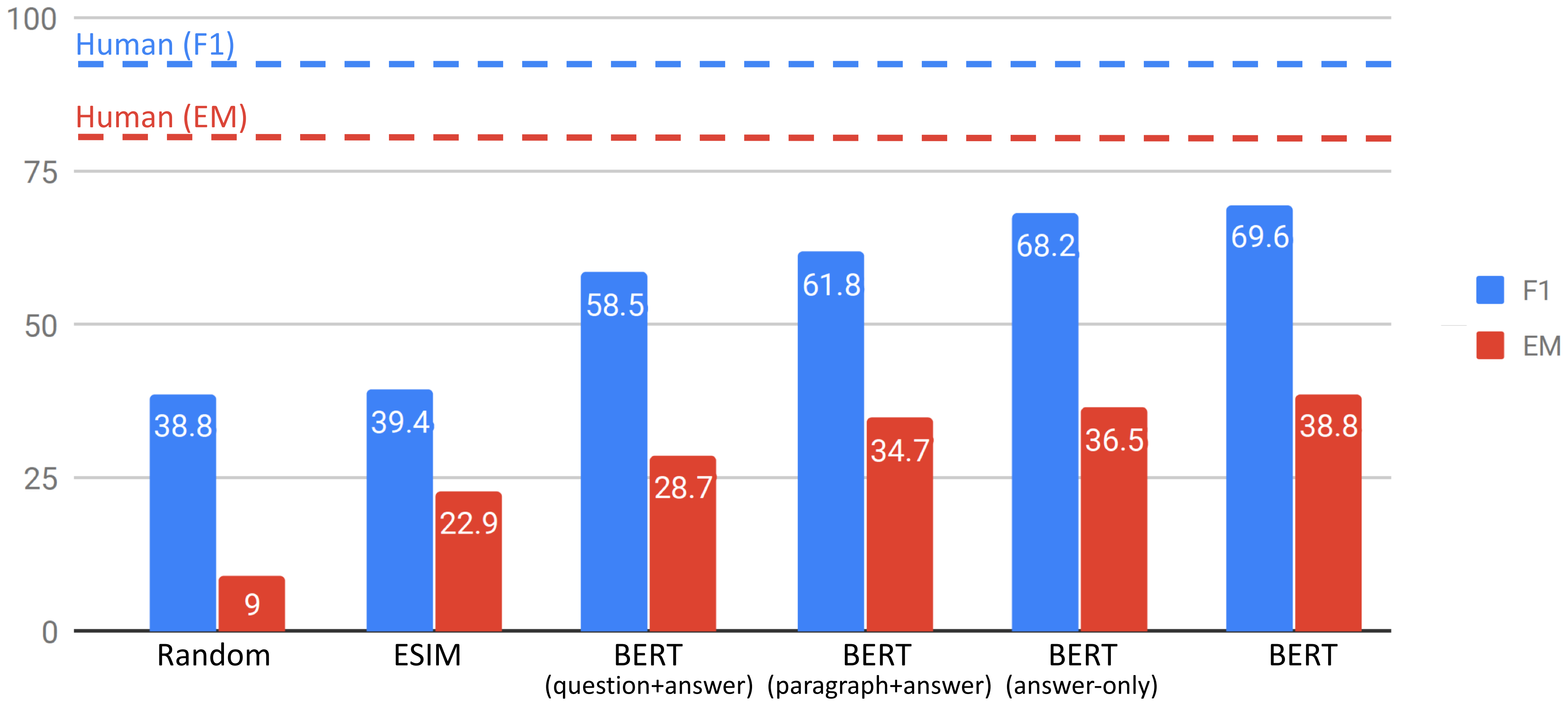


A systems gets credit only if, it gets **all** the candidates right.

Supervised-learning (LSTM)
[Chen et al, 2017]

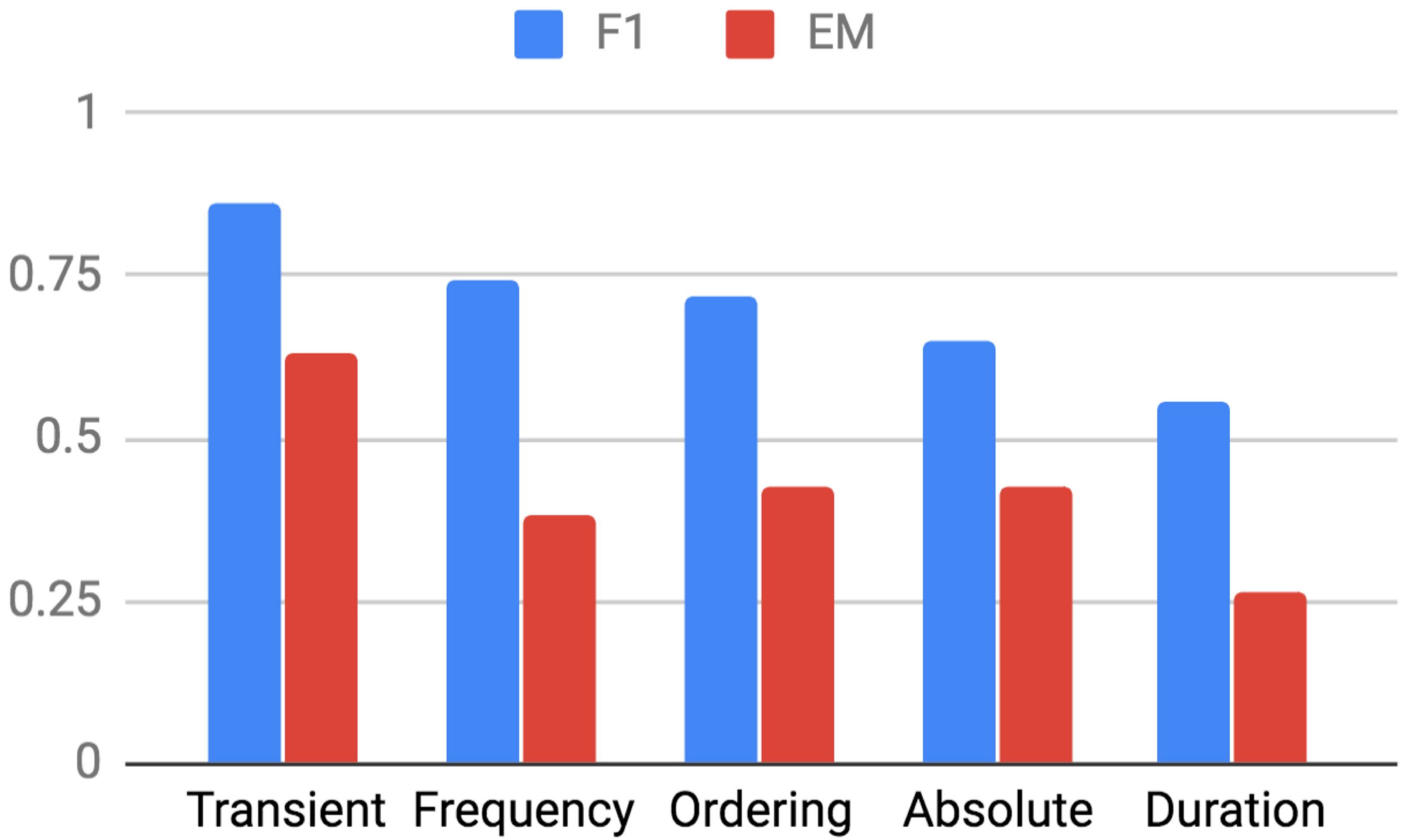
Pre-trained contextualized model
[Devlin et al, 2018]

Experimental results [ZKNR, under review]



[Devlin et al, 2018]

Experimental results [ZKNR, under review]



The State of Current Systems

[Liu & Singh, 2001]

[Devlin et al, 2018; Peters et al, 2018]

The State of Current Systems

- Explicit knowledge bases; e.g. ConceptNet [Liu & Singh, 2001]
  - Okay precision but low recall (coverage)
  - Suffer from brittleness
- Language models, soft representations [Devlin et al, 2018; Peters et al, 2018]
  - Can deal with certain associations
  - Suffer from precision issues

The State of Current Systems

- Explicit knowledge bases; e.g. ConceptNet [Liu & Singh, 2001]
  - Okay precision but low recall (coverage)
  - Suffer from brittleness
- Language models, soft representations [Devlin et al, 2018; Peters et al, 2018]
  - Can deal with certain associations
  - Suffer from precision issues

The State of Current Systems

- Explicit knowledge bases; e.g. ConceptNet [Liu & Singh, 2001]
  - Okay precision but low recall (coverage)
  - Suffer from brittleness
- Language models, soft representations [Devlin et al, 2018; Peters et al, 2018]
  - Can deal with certain associations
  - Suffer from precision issues

Scenario: He laid down on the chair and pawed at her as she ran in a circle under it.
Question: How long did she run in a circle?

The State of Current Systems

- Explicit knowledge bases; e.g. ConceptNet [Liu & Singh, 2001]
  - Okay precision but low recall (coverage)
  - Suffer from brittleness
- Language models, soft representations [Devlin et al, 2018; Peters et al, 2018]
  - Can deal with certain associations
  - Suffer from precision issues

Scenario: He laid down on the chair and pawed at her as she ran in a circle under it.
Question: How long did she run in a circle?

BERT selected as answer	Candidate answer
✗	weeks
✗	days
✓	minutes

The State of Current Systems

- Explicit knowledge bases; e.g. ConceptNet [Liu & Singh, 2001]
  - Okay precision but low recall (coverage)
  - Suffer from brittleness
- Language models, soft representations [Devlin et al, 2018; Peters et al, 2018]
  - Can deal with certain associations
  - Suffer from precision issues

Scenario: He laid down on the chair and pawed at her as she ran in a circle under it.
Question: How long did she run in a circle?

BERT selected as answer	Candidate answer
✗	weeks
✗	days
✓	minutes
✓	1 minutes
✓	28740000 minutes

Summary of This Section

- Understanding time is crucial aspect of NLU.
- A QA dataset of temporal commonsense questions.
- Evaluated systems and showing few angles they are missing.

Road Map

- **Part 2:** Moving the Peaks Higher: More Challenging Datasets
  - A QA Benchmark for Temporal Common-sense *[Submitted]*
  - A QA Benchmark for Reasoning on Multiple Sentences *[NAACL'18]*

vignettes

A Benchmark for Reasoning over Multiple Sentences [KCRUR'18]

vignettes

A Benchmark for Reasoning over Multiple Sentences [KCRUR'18]

"Multi-sentence" hypothesis:
Questions that require multiple sentences tend to be "hard".

vignettes

A Benchmark for Reasoning over Multiple Sentences [KCRUR'18]

- The need for creating "reasoning-forcing" challenges

"Multi-sentence" hypothesis:
Questions that require multiple sentences tend to be "hard".

- 4-step crowdsourcing
- From 8 domains (fiction, news, science, etc)
  - +10k questions
  - 50k candidate answers
  - +700 paragraphs

`https://cogcomp.org/multirc`

vignettes

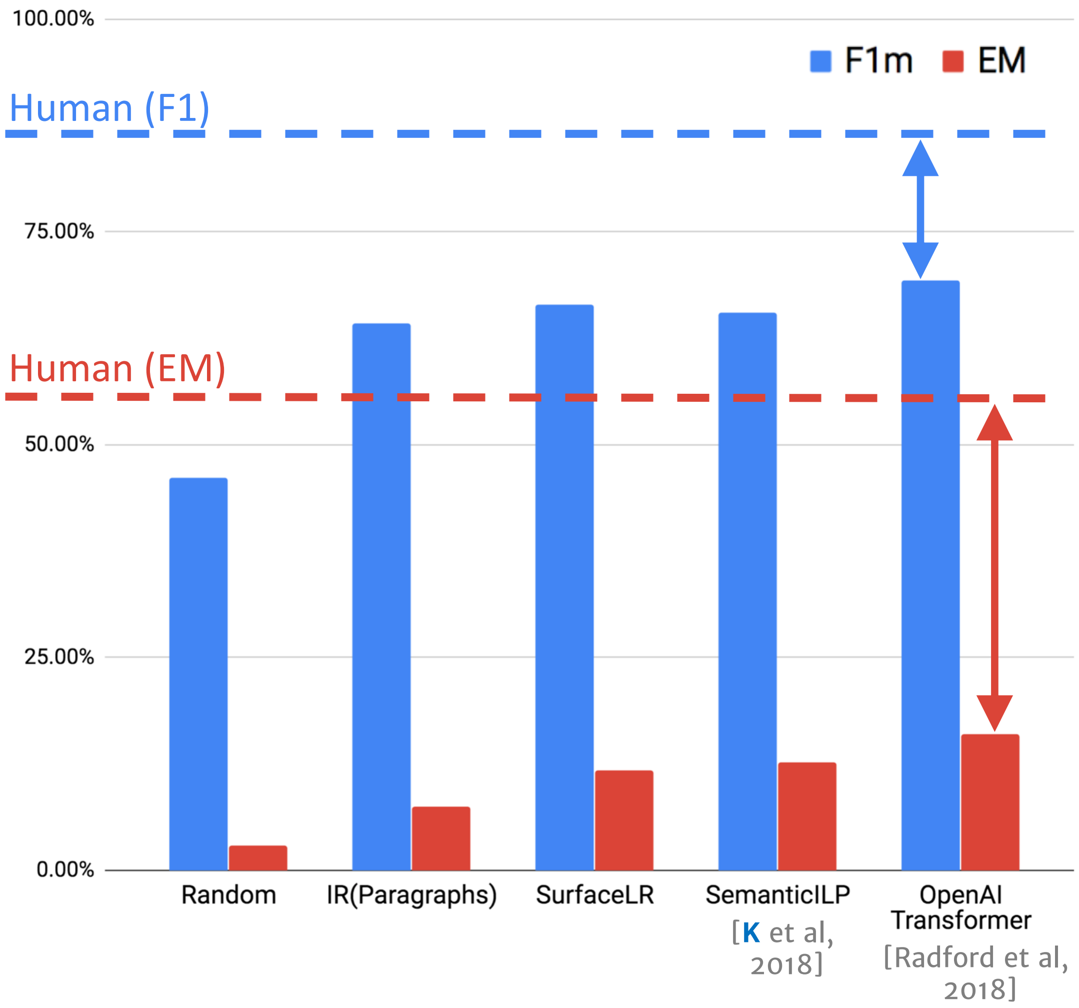
A Benchmark for Reasoning over Multiple Sentences [KCRUR'18]

- The need for creating "reasoning-forcing" challenges

"Multi-sentence" hypothesis:
Questions that require multiple sentences tend to be "hard".

- 4-step crowdsourcing
- From 8 domains (fiction, news, science, etc)
  - +10k questions
  - 50k candidate answers
  - +700 paragraphs

https://cogcomp.org/multirc



Road Map

- **Part 3:** Formal Study of Reasoning in Natural Language
  - Capabilities and Limitations of Reasoning in Natural Language *[In submission]*

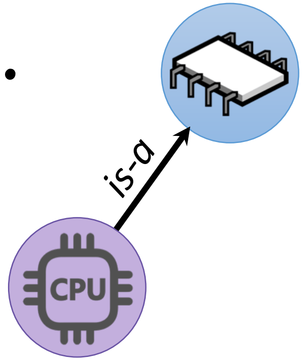
A Formal Study of NL Reasoning: Overview

A Formal Study of NL Reasoning: Overview

- We provide a formalized study of reasoning.
- Requires assumptions about "knowledge" and "reasoning".
  - Information represented as **graphs** (nodes and semantic relations).
    - Any other structure can be thought of an explicit or implicit graph.
  - Incorporate properties like *variability*, *ambiguity*, etc.
  - *Reasoning: the* operation that combines chunks of information to make a conclusion.
- Distinguish **successful** and **failed** reasoning.

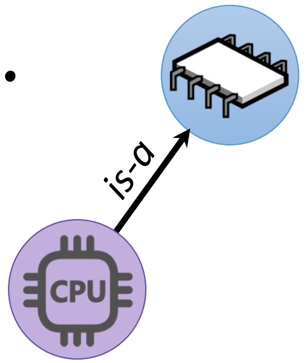
A Formal Study of NL Reasoning: Overview

- We provide a formalized study of reasoning.
- Requires assumptions about "knowledge" and "reasoning".
  - Information represented as **graphs** (nodes and semantic relations).
    - Any other structure can be thought of an explicit or implicit graph.
  - Incorporate properties like *variability*, *ambiguity*, etc.
  - *Reasoning: the* operation that combines chunks of information to make a conclusion.
- Distinguish **successful** and **failed** reasoning.



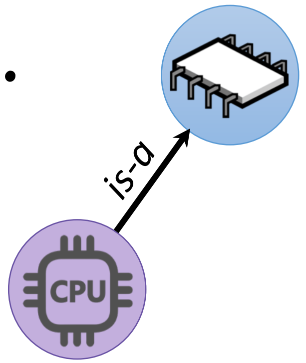
A Formal Study of NL Reasoning: Overview

- We provide a formalized study of reasoning.
- Requires assumptions about "knowledge" and "reasoning".
  - Information represented as **graphs** (nodes and semantic relations).
    - Any other structure can be thought of an explicit or implicit graph.
  - Incorporate properties like *variability*, *ambiguity*, etc.
  - *Reasoning: the* operation that combines chunks of information to make a conclusion.
- Distinguish **successful** and **failed** reasoning.



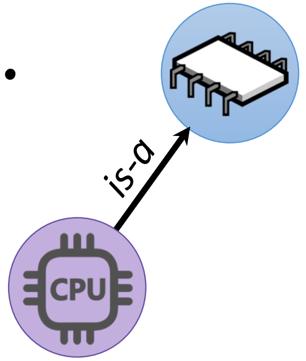
A Formal Study of NL Reasoning: Overview

- We provide a formalized study of reasoning.
- Requires assumptions about "knowledge" and "reasoning".
  - Information represented as **graphs** (nodes and semantic relations).
    - Any other structure can be thought of an explicit or implicit graph.
  - Incorporate properties like *variability*, *ambiguity*, etc.
  - *Reasoning: the* operation that combines chunks of information to make a conclusion.
- Distinguish **successful** and **failed** reasoning.



A Formal Study of NL Reasoning: Overview

- We provide a formalized study of reasoning.
- Requires assumptions about "knowledge" and "reasoning".
  - Information represented as **graphs** (nodes and semantic relations).
    - Any other structure can be thought of an explicit or implicit graph.
  - Incorporate properties like *variability*, *ambiguity*, etc.
  - *Reasoning: the* operation that combines chunks of information to make a conclusion.
- Distinguish **successful** and **failed** reasoning.



What We Do Not Say

What We Do Not Say

- Not making claims about:
  - How “reasoning” should be defined.
  - How “knowledge” should be represented.
  - How systems should be designed.
- Theoretical results based on assumptions (“no free lunch”).
  - which may or may not stand the test of time.

What We Do Not Say

- Not making claims about:
  - How “reasoning” should be defined.
  - How “knowledge” should be represented.
  - How systems should be designed.
- Theoretical results based on assumptions (“no free lunch”).
  - which may or may not stand the test of time.

Tale of Two Spaces

Meaning Graph

- *Conceptualization*
- *No ambiguity*
- *No variability*
- *No missing relations*

Tale of Two Spaces

Utterance

Meaning Graph

- *Conceptualization*
- *No ambiguity*
- *No variability*
- *No missing relations*

Tale of Two Spaces

Symbol Graph

- *Physical world*
- *Ambiguity*
- *Variability*
- *Missing relations*

Utterance

Meaning Graph

- *Conceptualization*
- *No ambiguity*
- *No variability*
- *No missing relations*

WIKIDATA

Tale of Two Spaces

Symbol Graph

- *Physical world*
- *Ambiguity*
- *Variability*
- *Missing relations*

Utterance

Meaning Graph

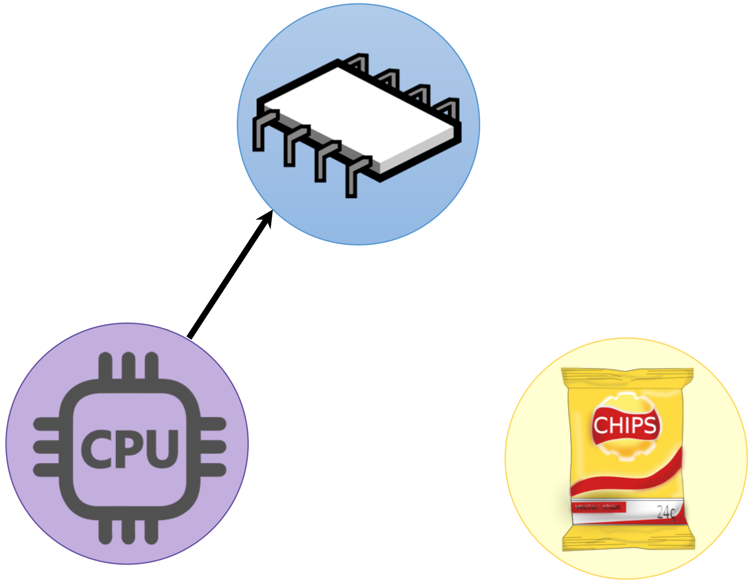
- *Conceptualization*
- *No ambiguity*
- *No variability*
- *No missing relations*

WIKIDATA

Formalizing Symbol and Meaning Space

Symbol Graph

Meaning Graph

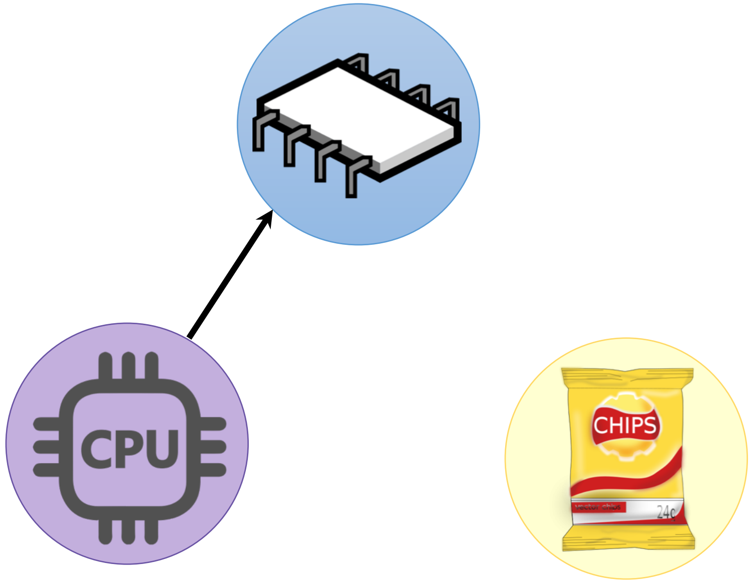


Formalizing Symbol and Meaning Space

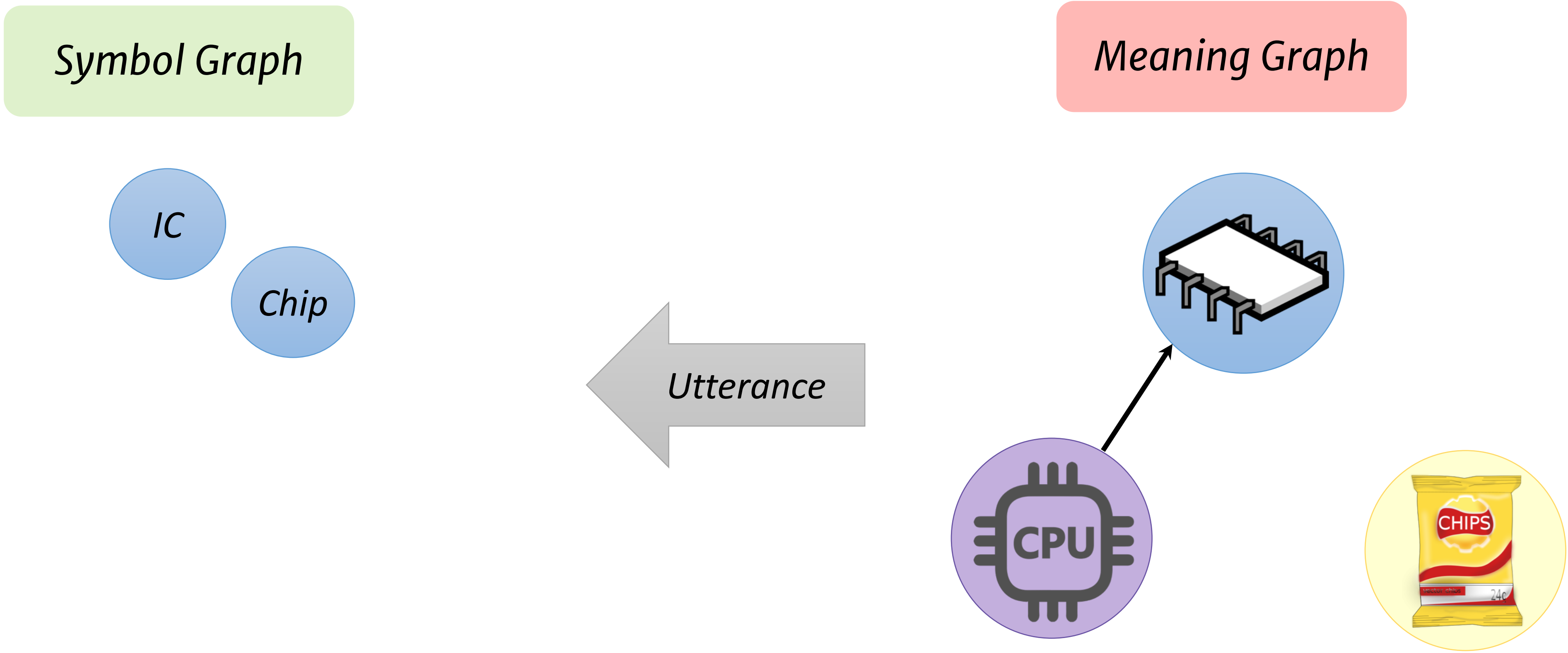
Symbol Graph

Meaning Graph

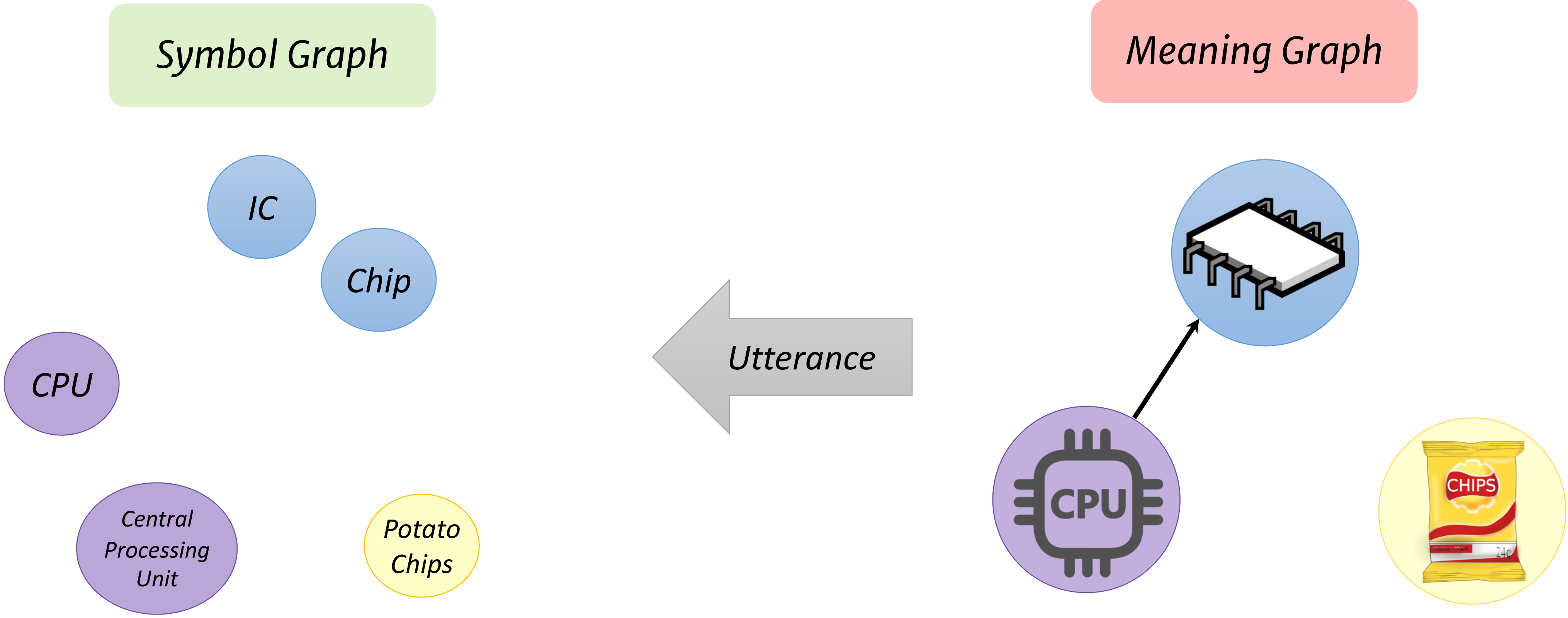
Utterance



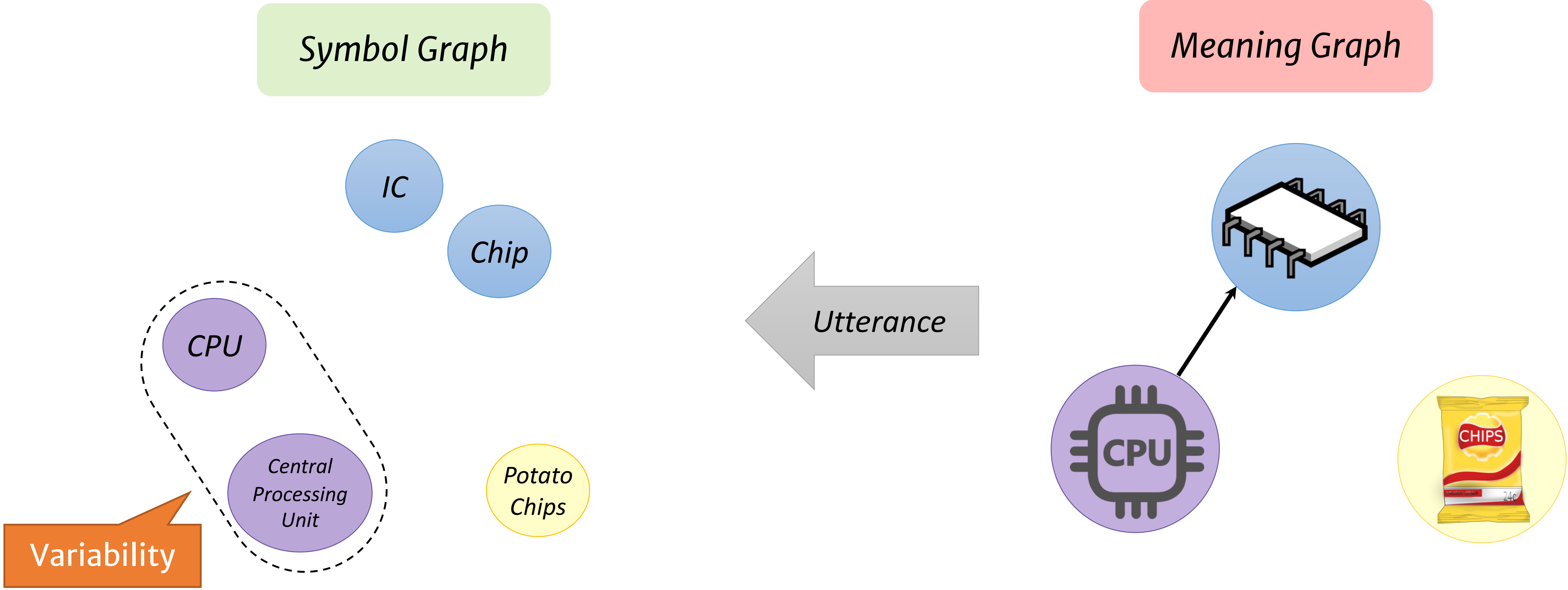
Formalizing Symbol and Meaning Space



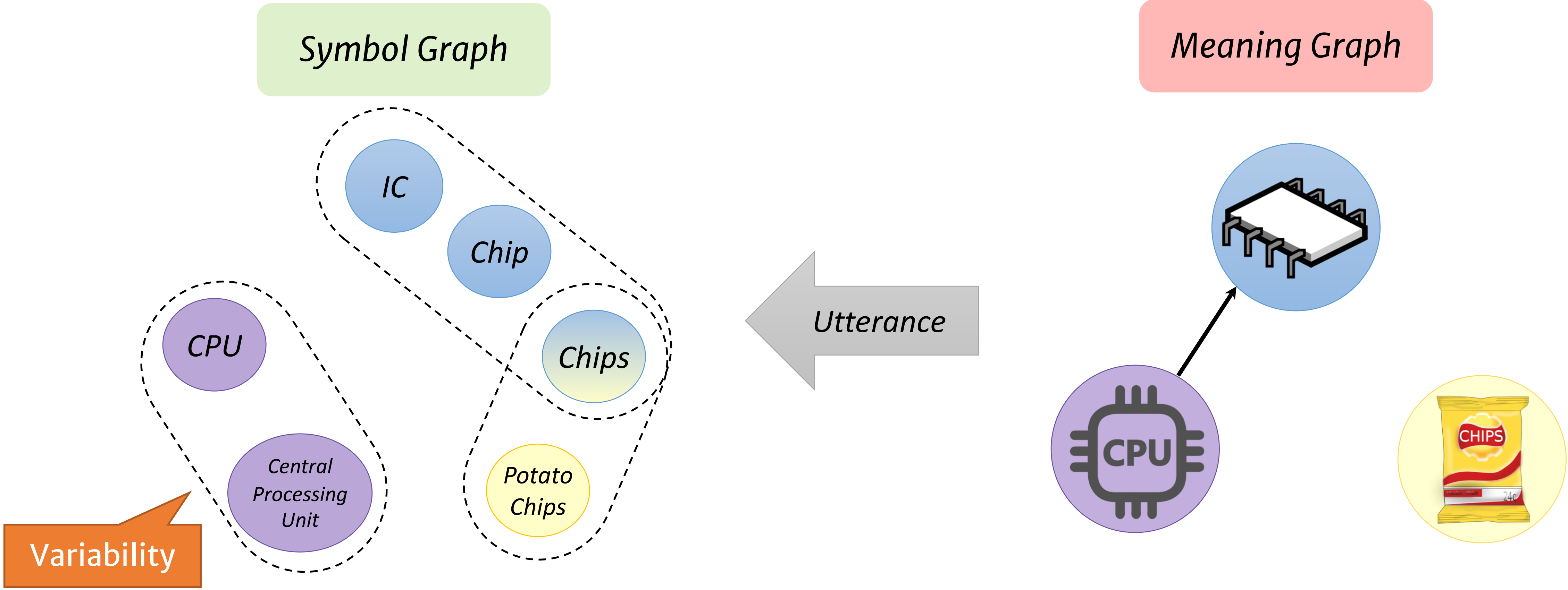
Formalizing Symbol and Meaning Space



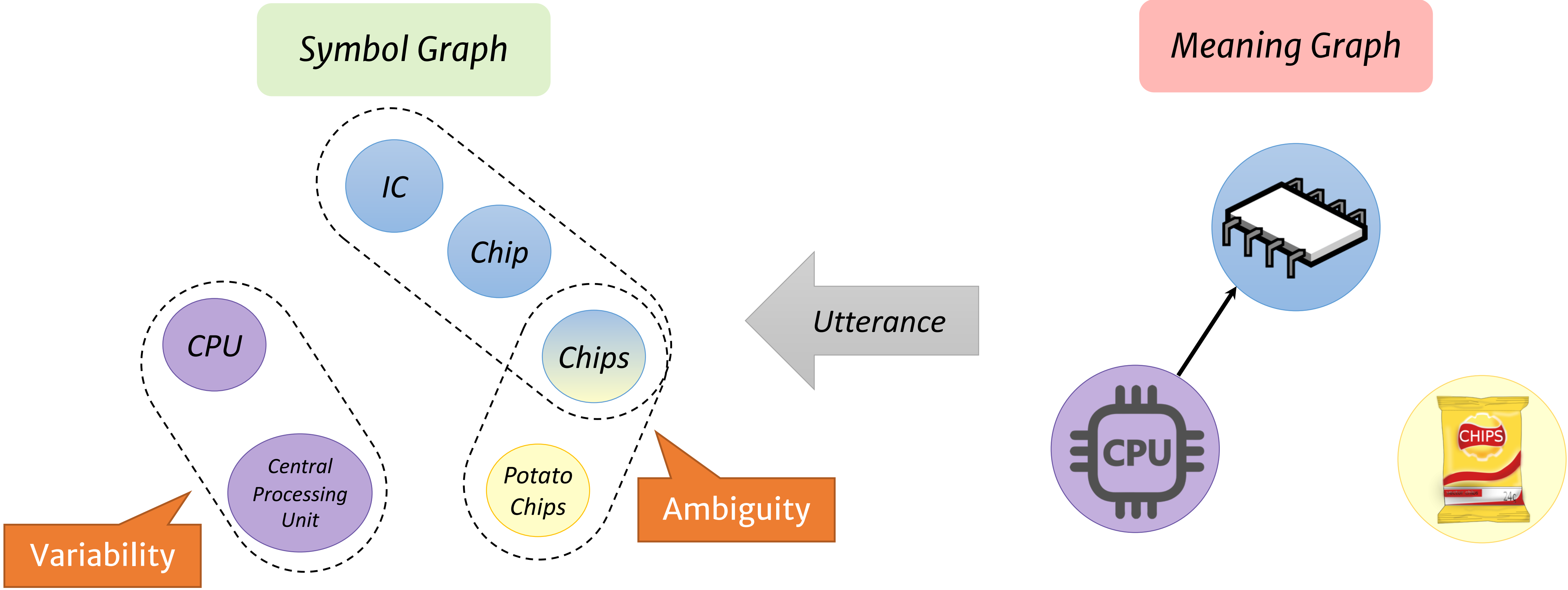
Formalizing Symbol and Meaning Space



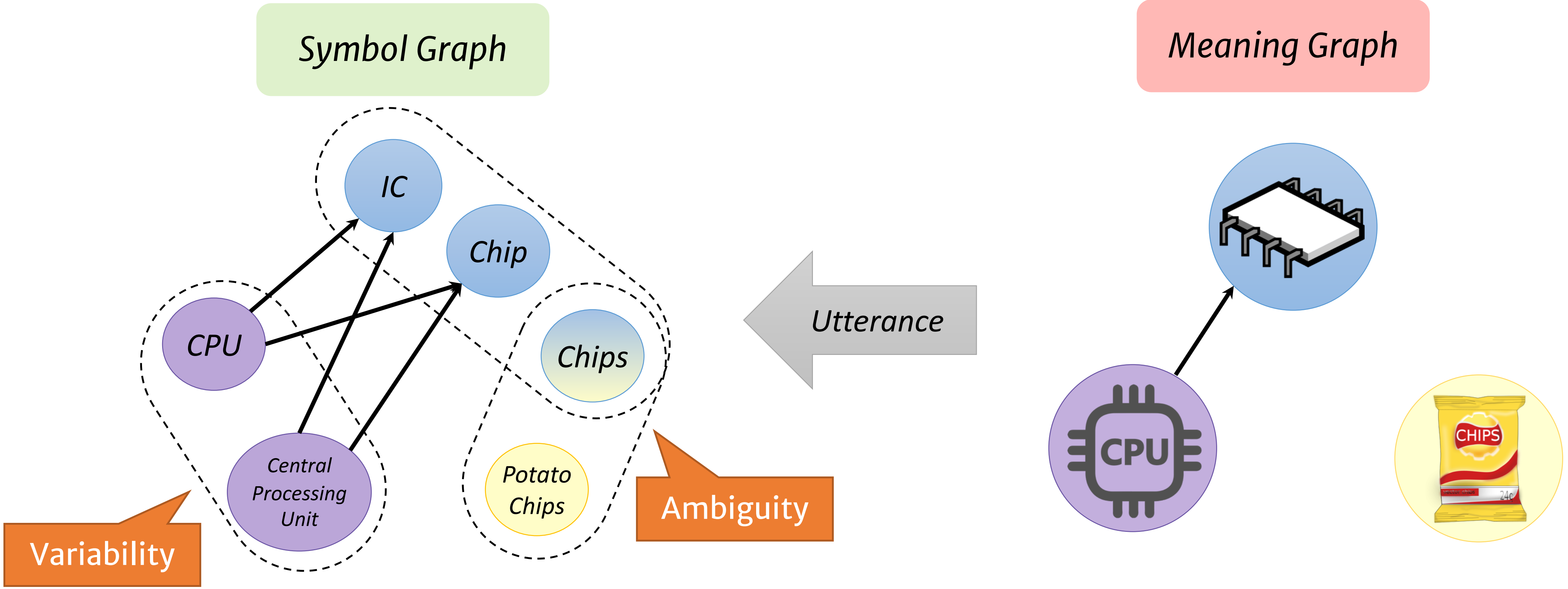
Formalizing Symbol and Meaning Space



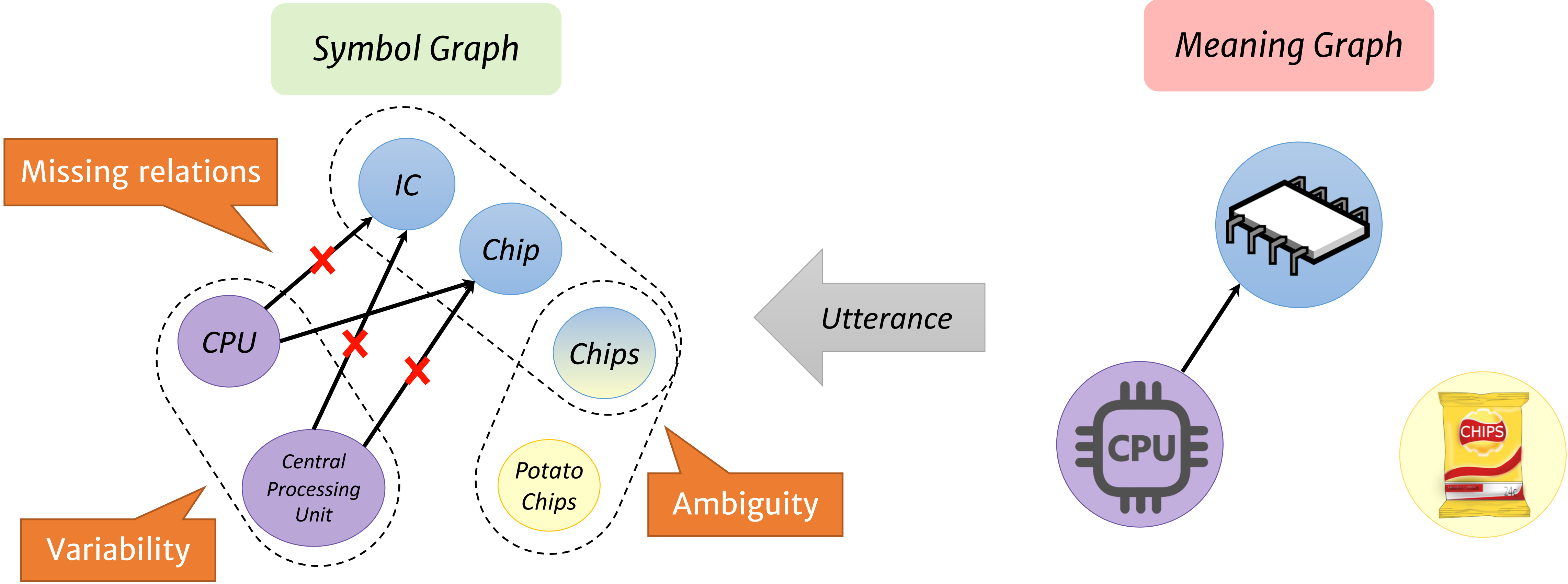
Formalizing Symbol and Meaning Space



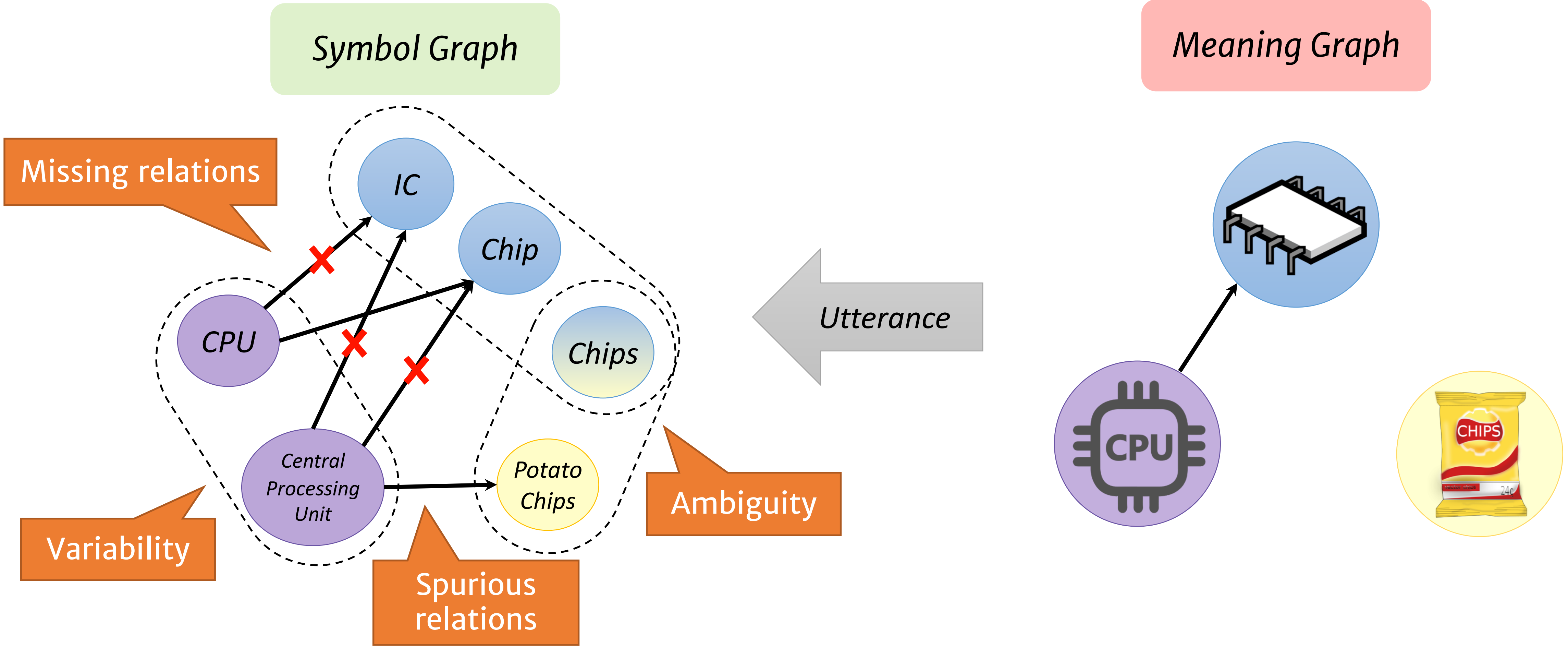
Formalizing Symbol and Meaning Space



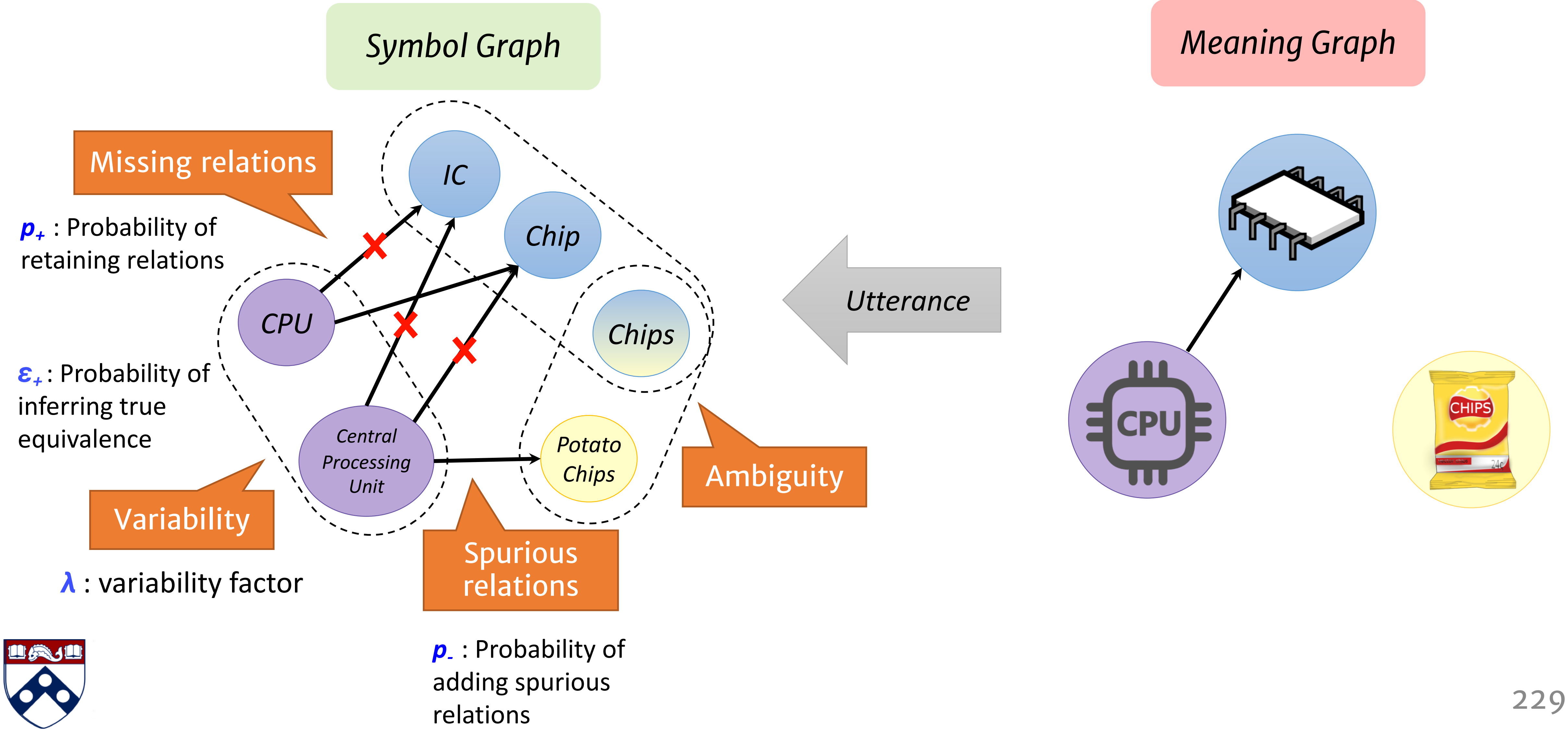
Formalizing Symbol and Meaning Space



Formalizing Symbol and Meaning Space



Formalizing Symbol and Meaning Space



Reasoning by Combining Local Information

- Reasoning itself is hard to define.
- Class of reasoning which functions by combining local information ("*multi-hop*")

Symbol Graph

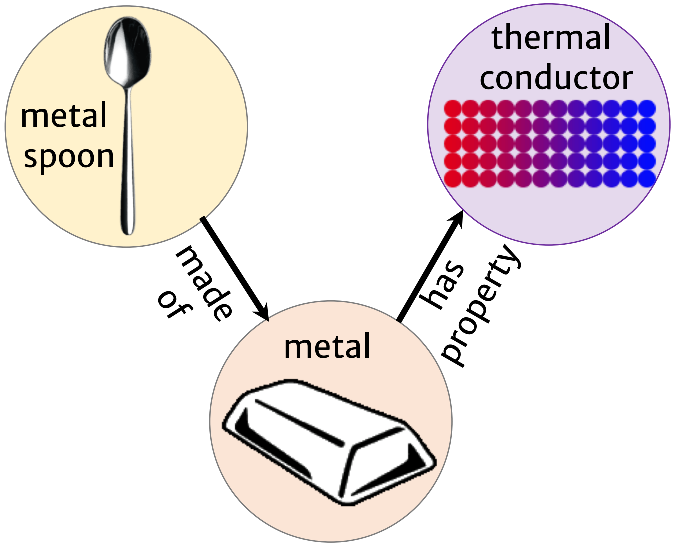
Meaning Graph

Reasoning by Combining Local Information

- Reasoning itself is hard to define.
- Class of reasoning which functions by combining local information ("*multi-hop*")

Symbol Graph

Meaning Graph

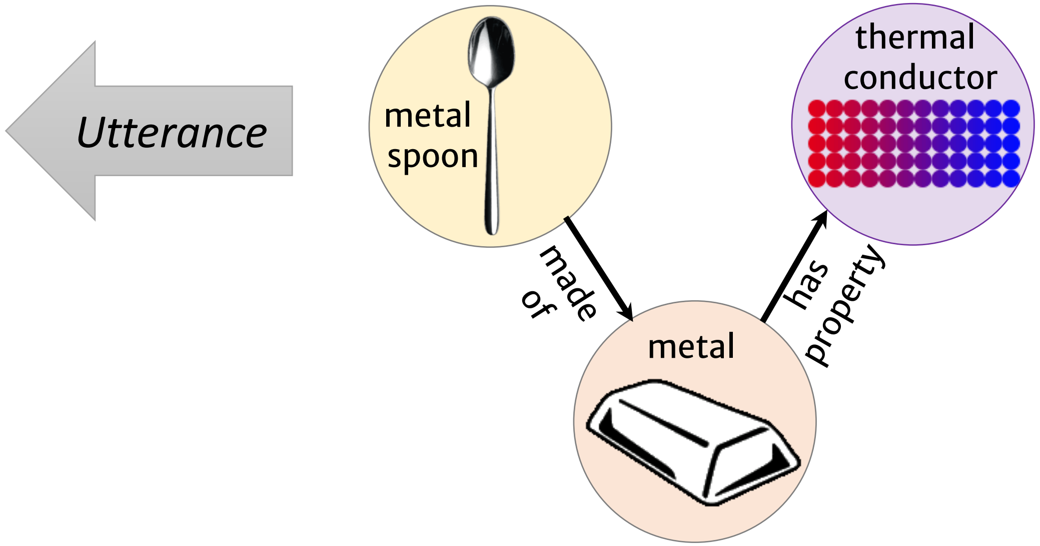


Reasoning by Combining Local Information

- Reasoning itself is hard to define.
- Class of reasoning which functions by combining local information (“*multi-hop*”)

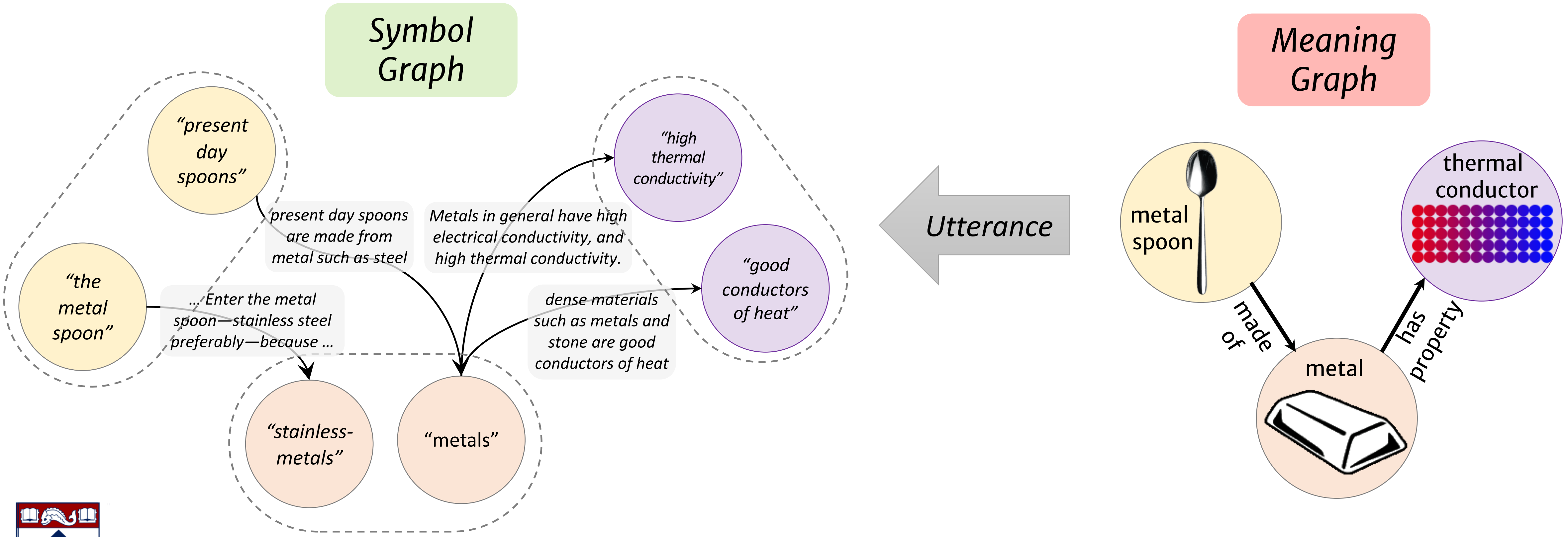
Symbol Graph

Meaning Graph



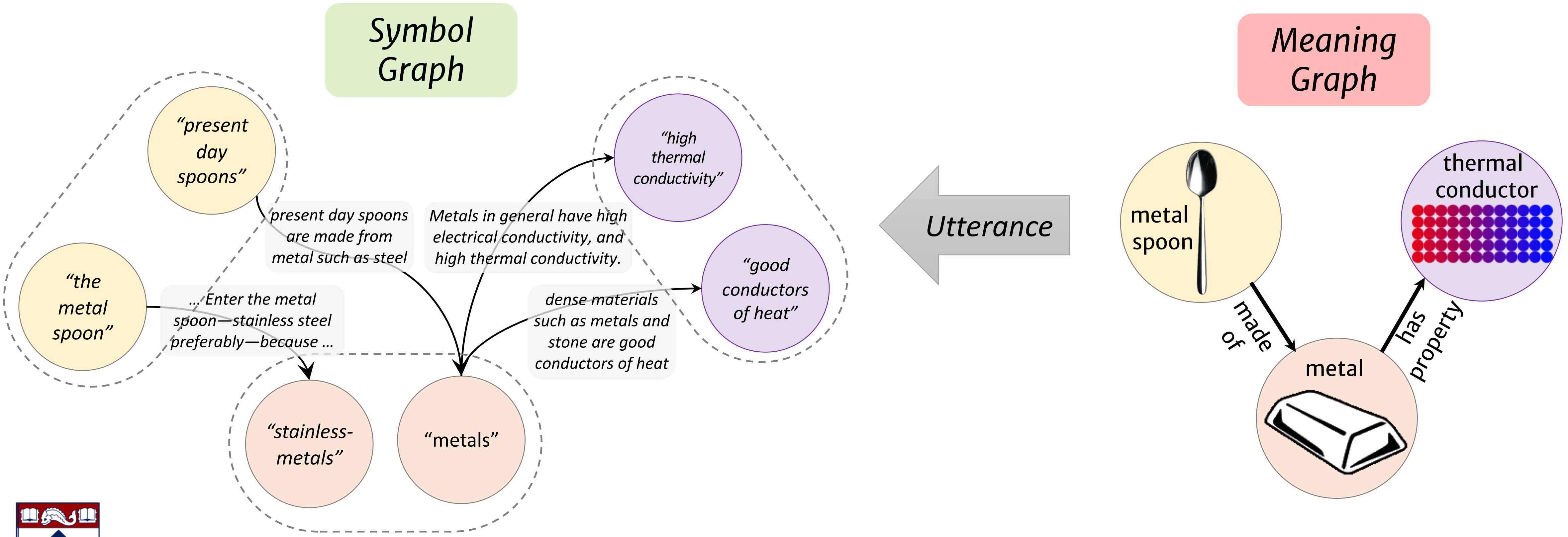
Reasoning by Combining Local Information

- Reasoning itself is hard to define.
- Class of reasoning which functions by combining local information (“*multi-hop*”)



Reasoning by Combining Local Information

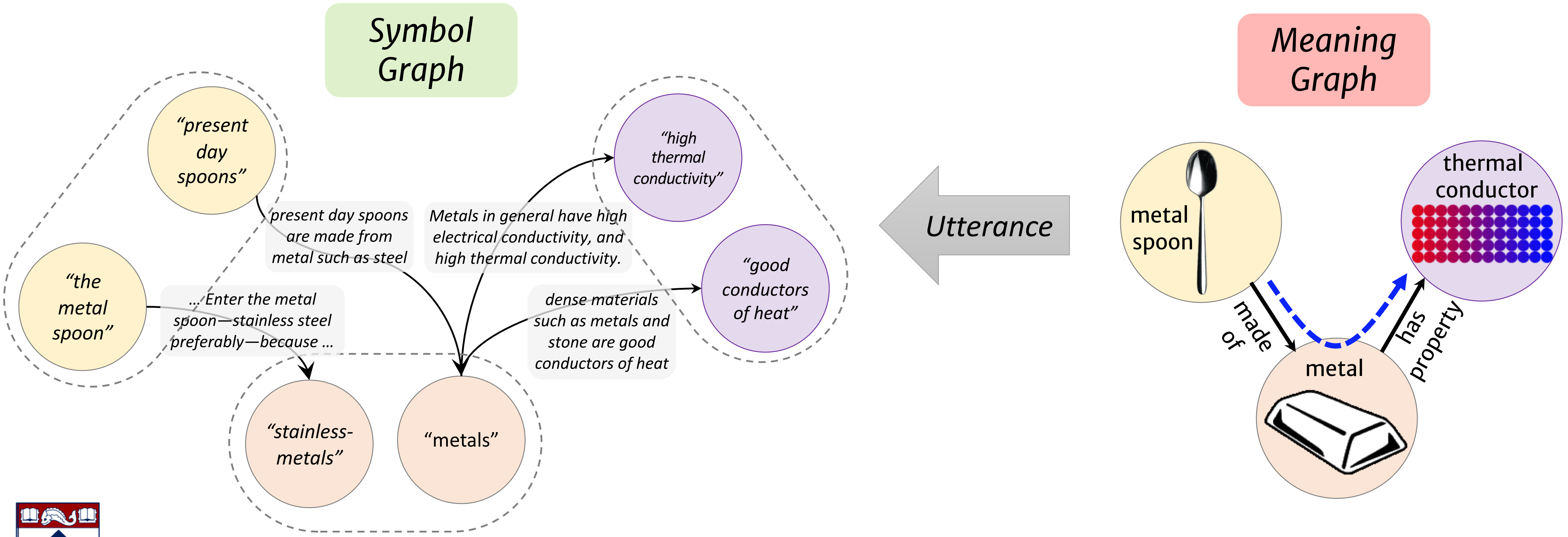
- Reasoning itself is hard to define.
- Class of reasoning which functions by combining local information (“*multi-hop*”)



Q: has-property(metal-spoon, thermal-conductor)

Reasoning by Combining Local Information

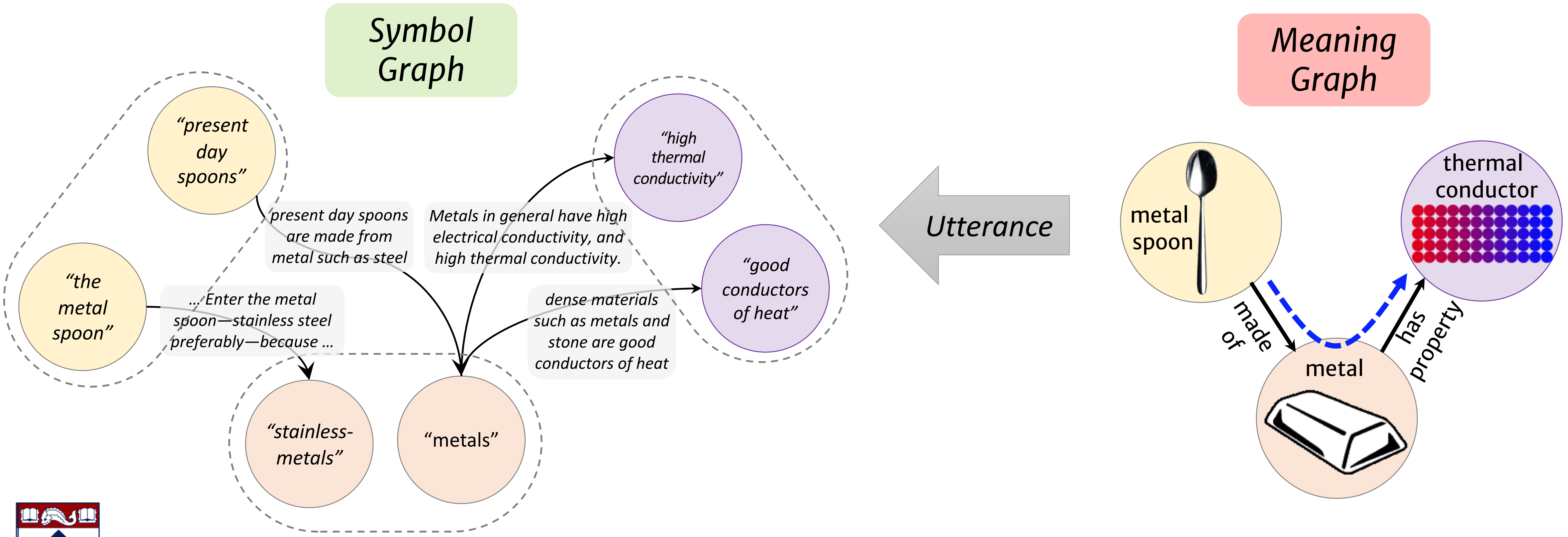
- Reasoning itself is hard to define.
- Class of reasoning which functions by combining local information (“*multi-hop*”)



Q: has-property(metal-spoon, thermal-conductor)

Reasoning by Combining Local Information

- Reasoning itself is hard to define.
- Class of reasoning which functions by combining local information ("*multi-hop*")

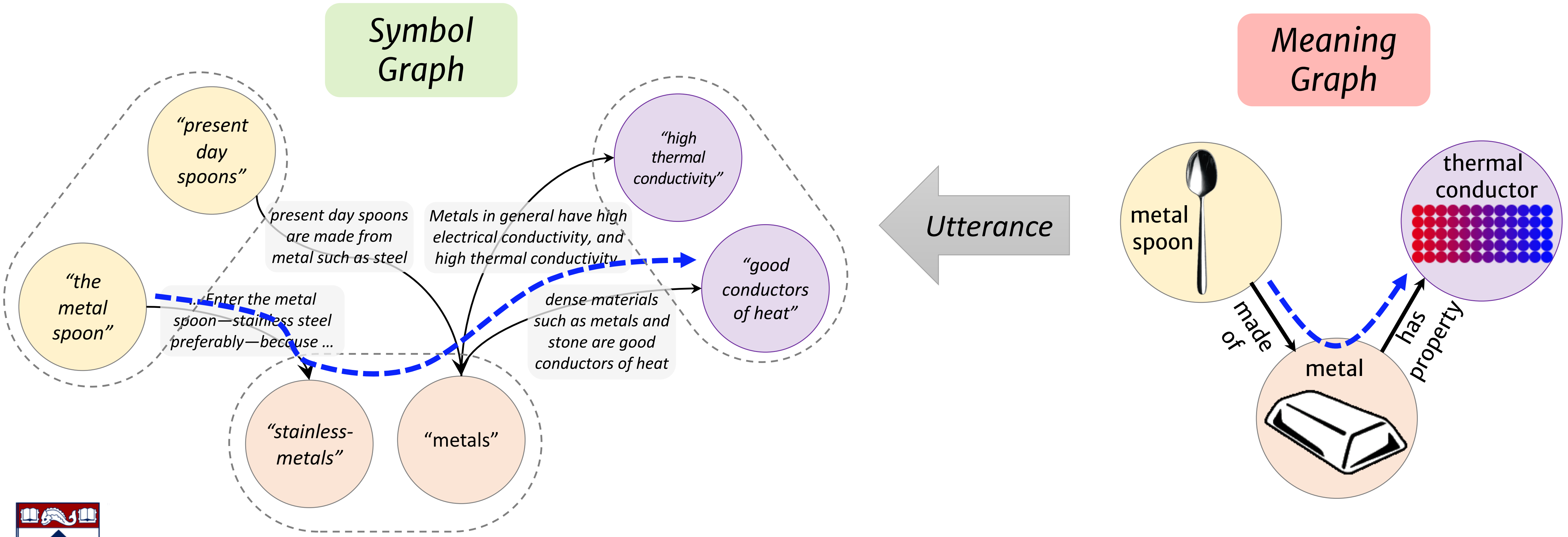


Q: "*is a metal spoon a good conductor of heat?*"

Q: has-property(metal-spoon, thermal-conductor)

Reasoning by Combining Local Information

- Reasoning itself is hard to define.
- Class of reasoning which functions by combining local information ("*multi-hop*")



Q: "*is a metal spoon a good conductor of heat?*"

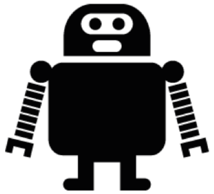
Q: has-property(metal-spoon, thermal-conductor)

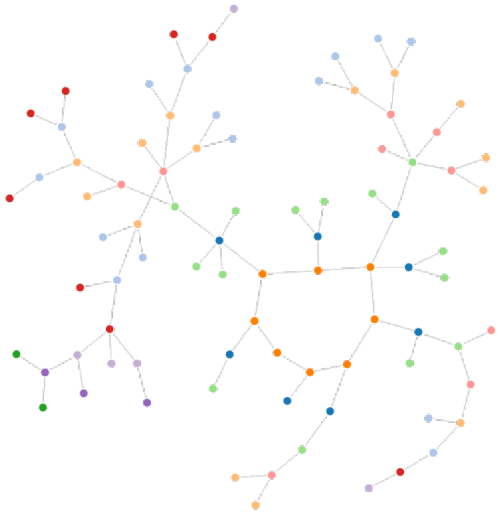
The Inference Problem

- "Inferring" connectivity in the (hidden) meaning graph
    - Given observations (a symbol graph)

The Inference Problem

- "Inferring" connectivity in the (hidden) meaning graph
  - Given observations (a symbol graph)

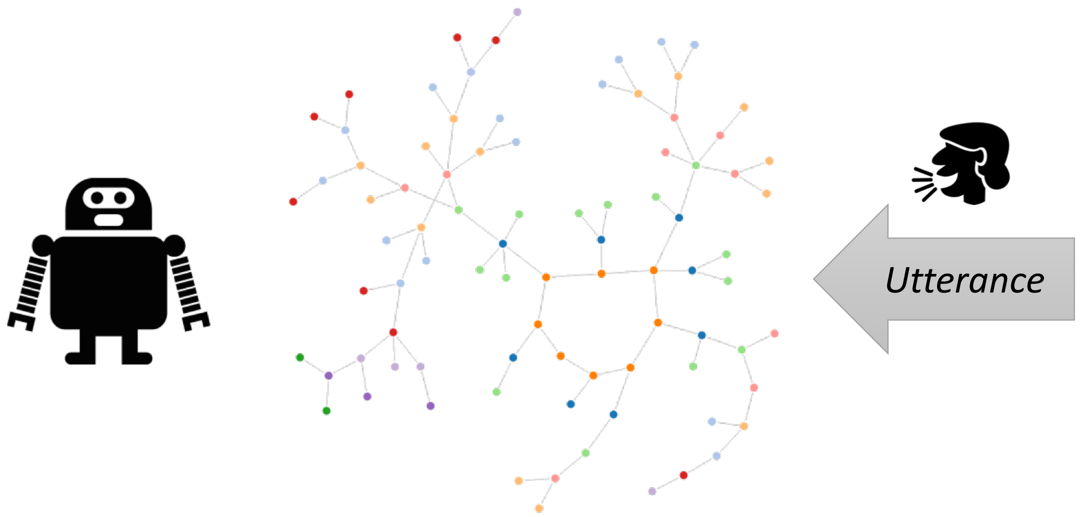




Symbol graph

The Inference Problem

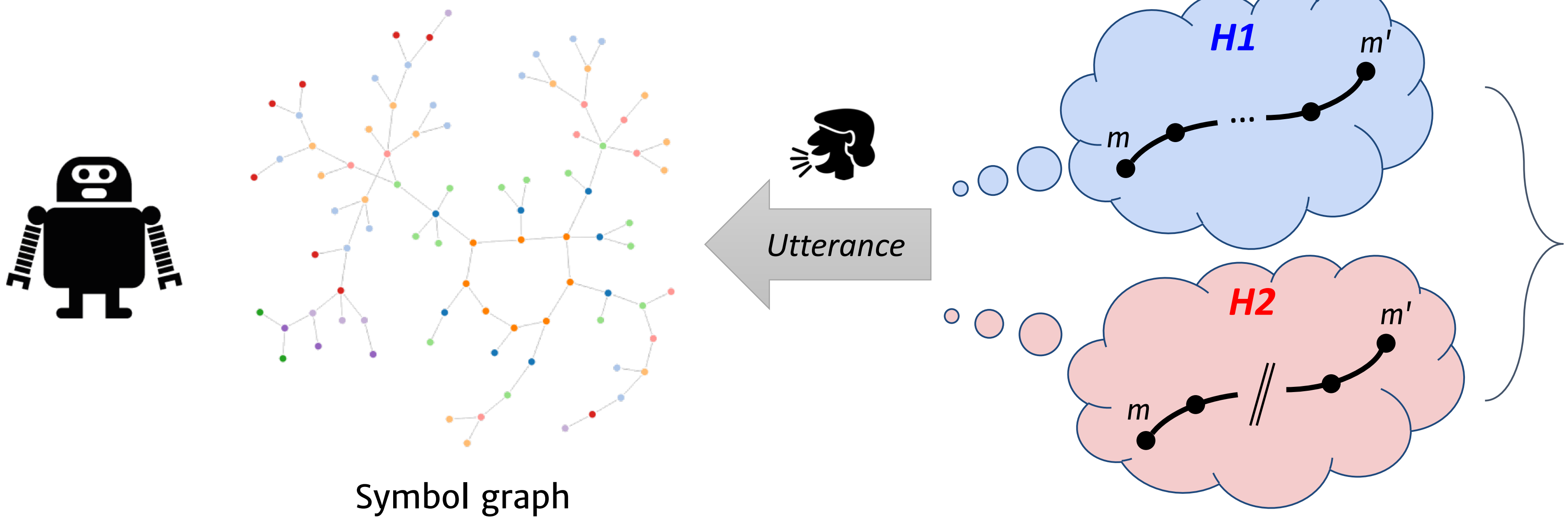
- "Inferring" connectivity in the (hidden) meaning graph
  - Given observations (a symbol graph)



Symbol graph

The Inference Problem

- “Inferring” connectivity in the (hidden) meaning graph
    - Given observations (a symbol graph)



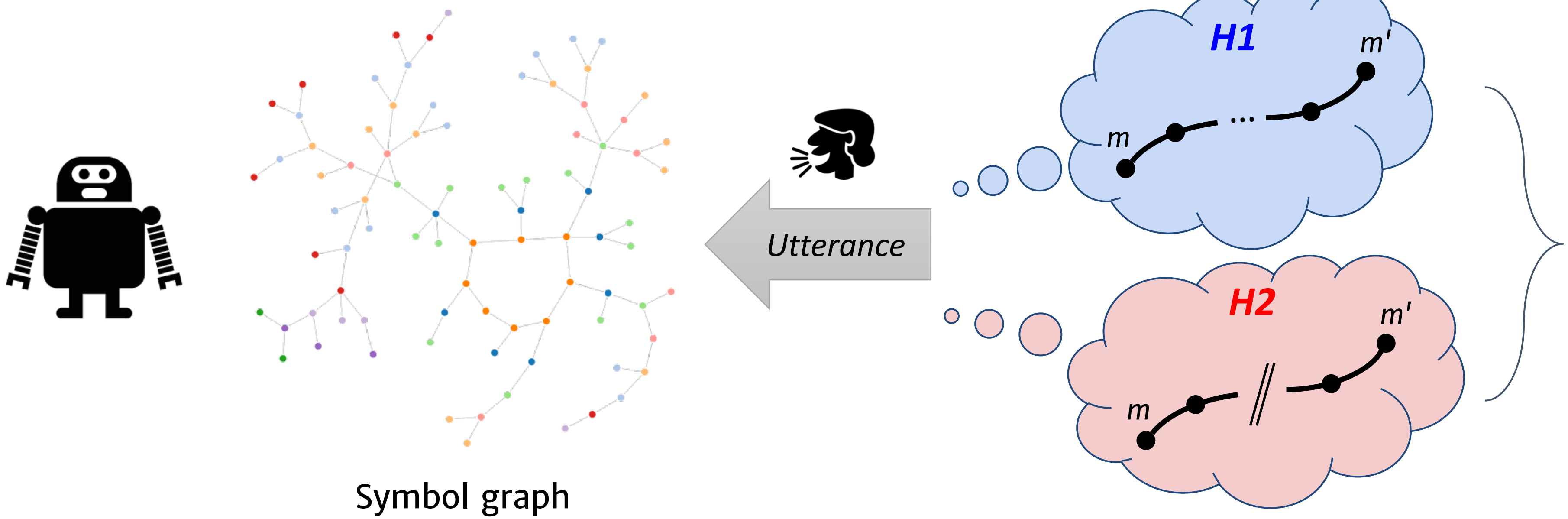
Symbol graph

Meaning graphs

one of the hypotheses require **d**-step connectivity

The Inference Problem

- "Inferring" connectivity in the (hidden) meaning graph
  - Given observations (a symbol graph)



Symbol graph

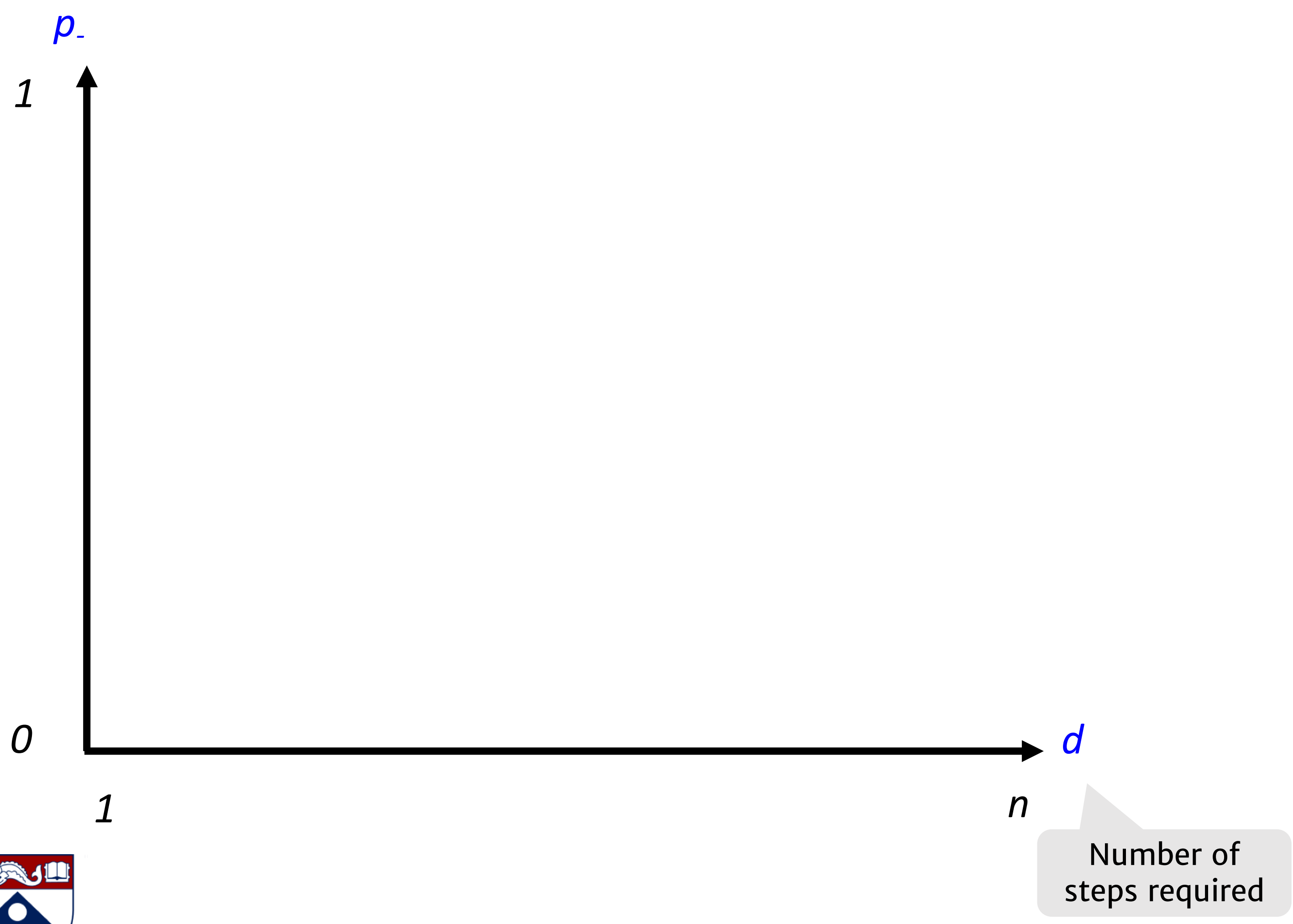
Meaning graphs

one of the hypotheses require **d**-step connectivity

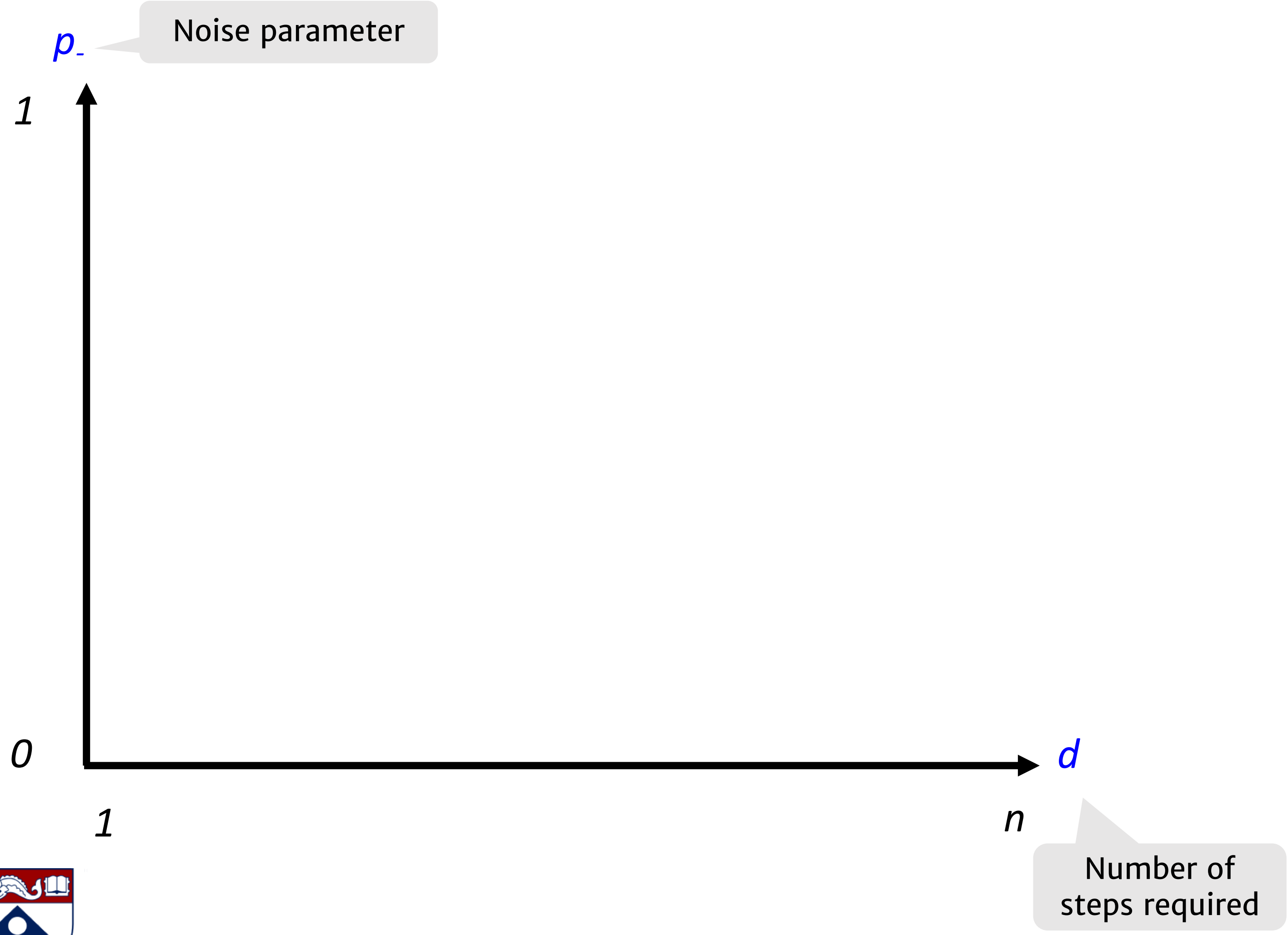
Goal: *Infer the connectivity of two given nodes (in the* ***unseen*** *meaning graph), given observations in the symbol graph.*

Results: Big picture [KSKSR, in submission]

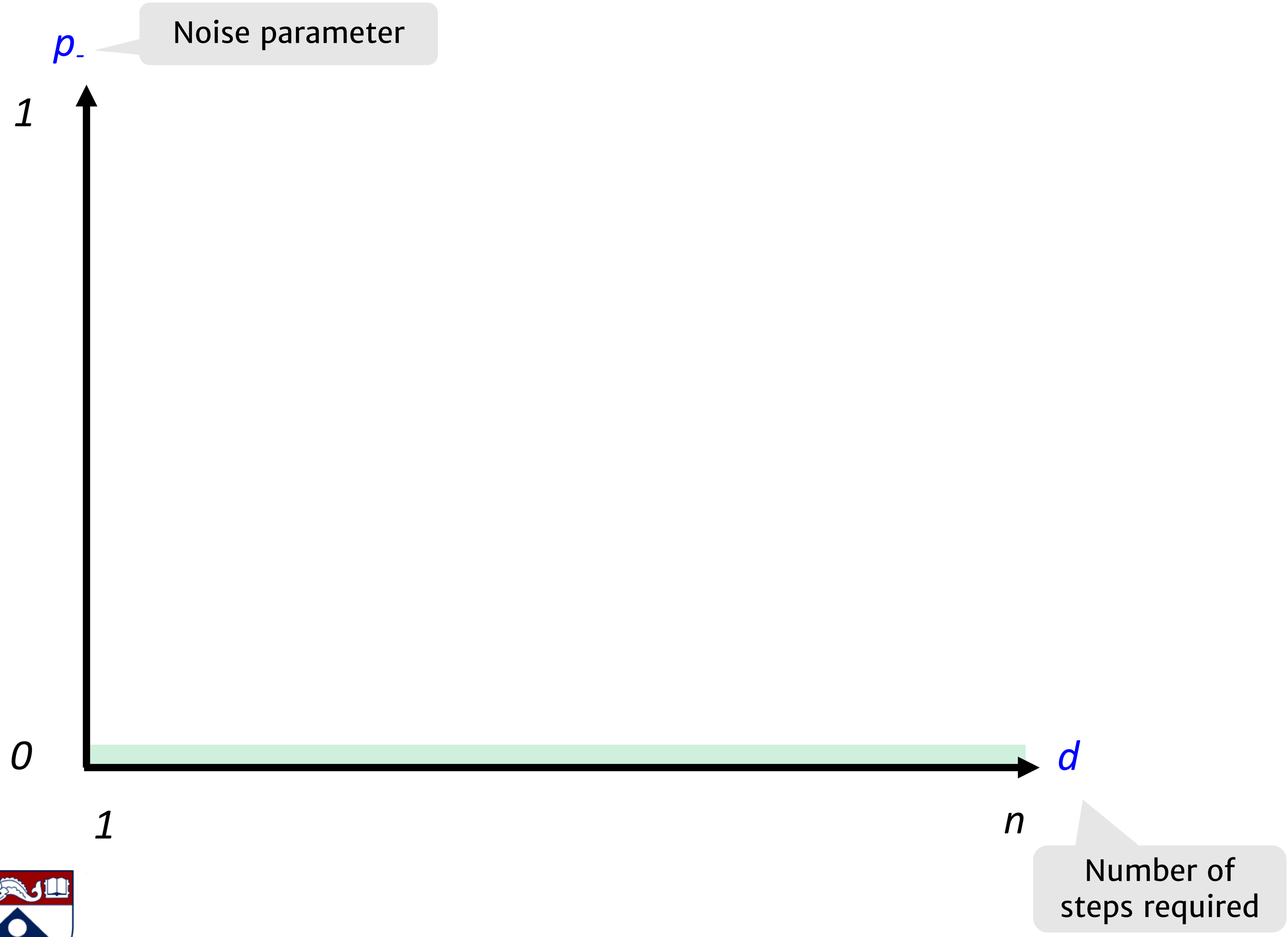
Results: Big picture [KSKSR, in submission]



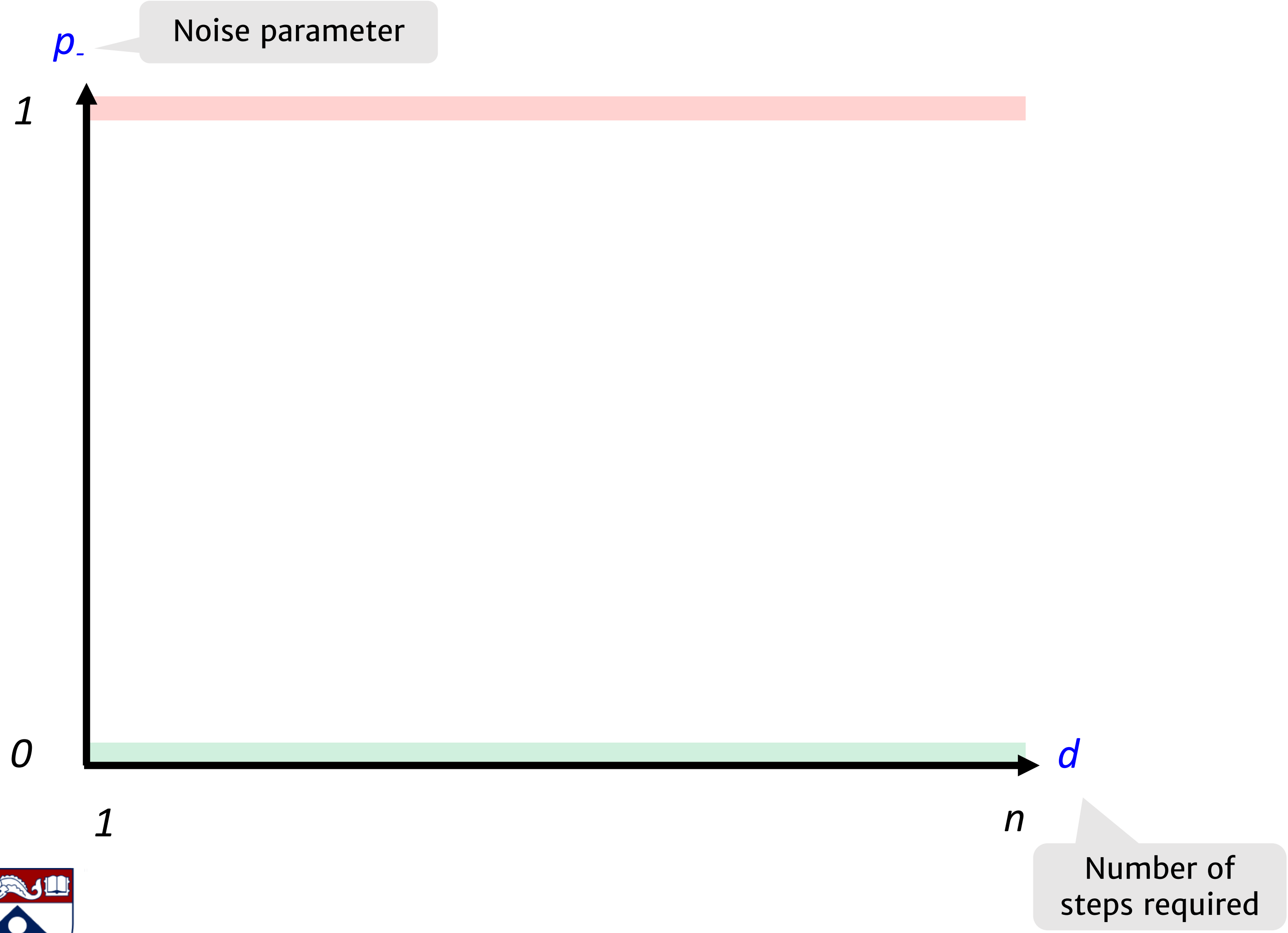
Results: Big picture [KSKSR, in submission]



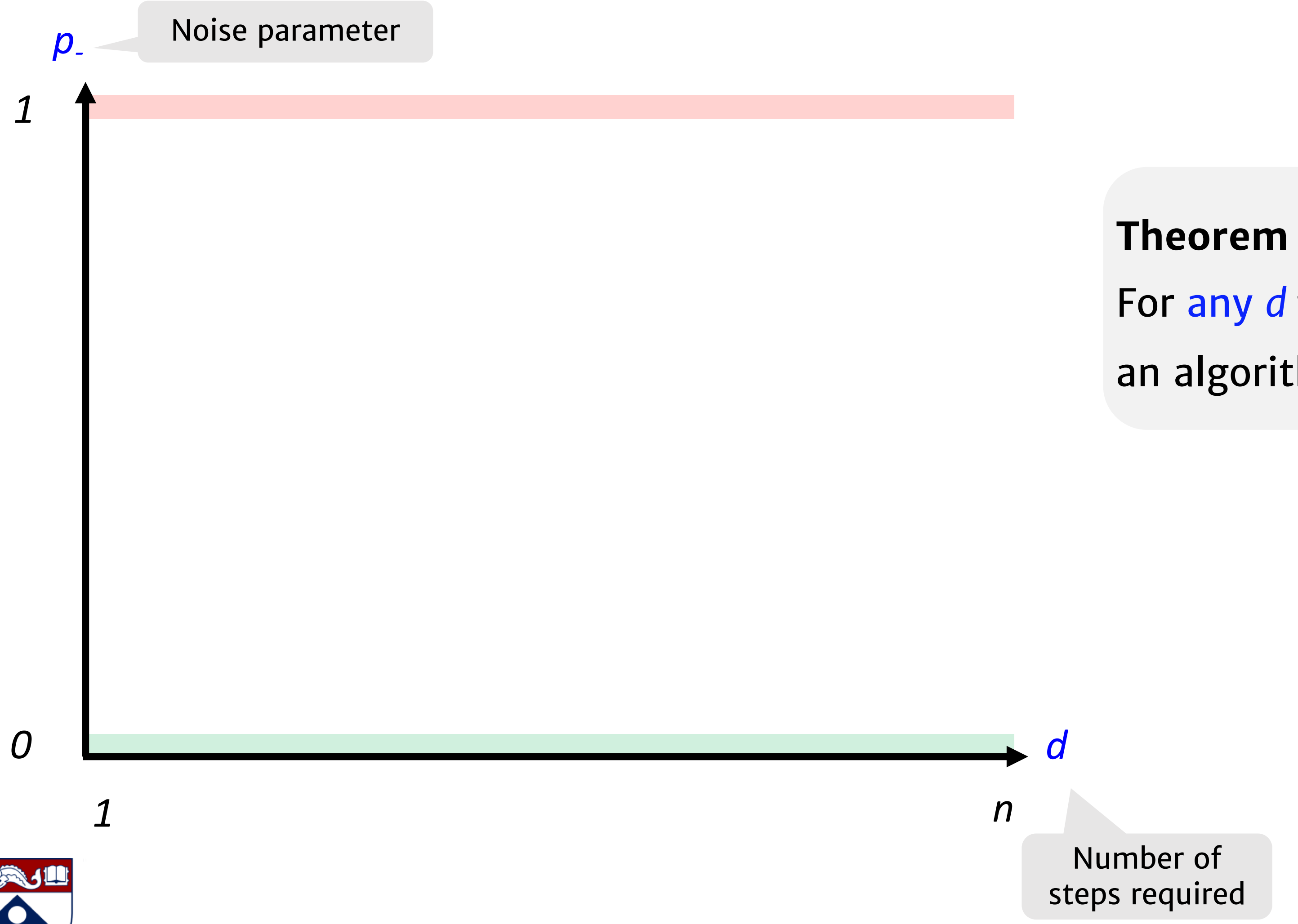
Results: Big picture [KSKSR, in submission]



Results: Big picture [KSKSR, in submission]



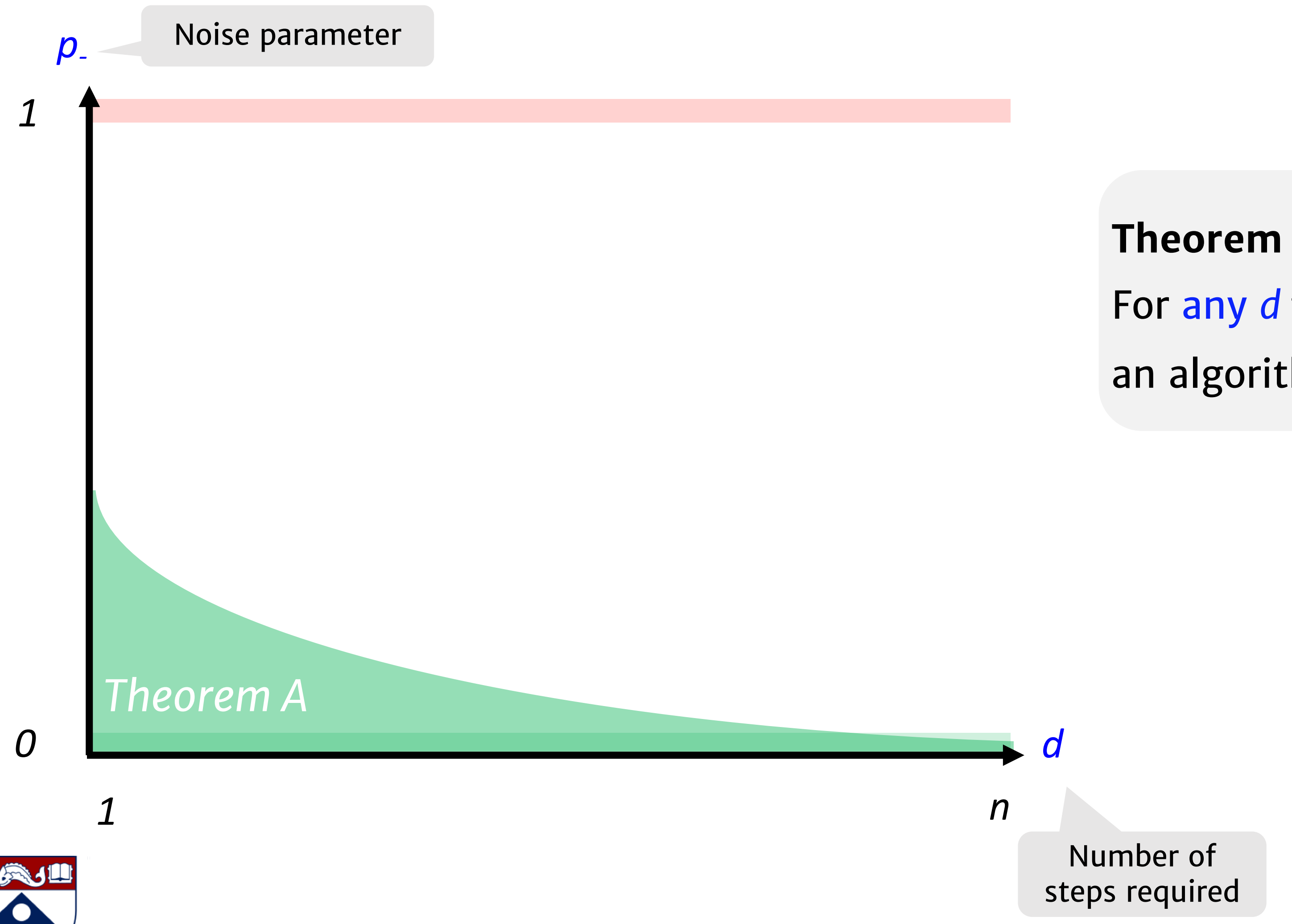
Results: Big picture [KSKSR, in submission]



Theorem A (informal)

For any d there is a choice of parameters such that an algorithm can confidently distinguish H_1 and H_2.

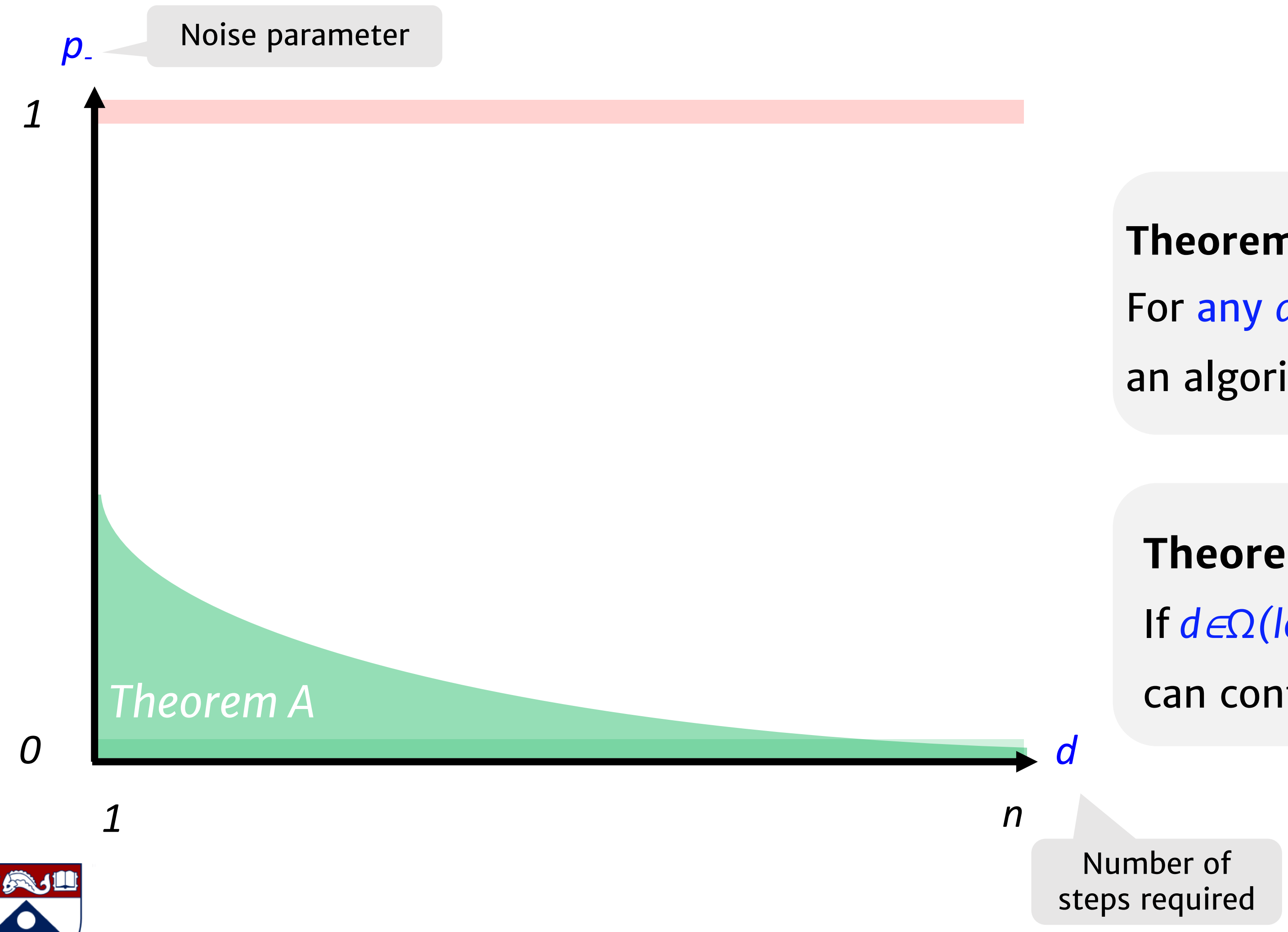
Results: Big picture [KSKSR, in submission]



Theorem A (informal)

For any d there is a choice of parameters such that an algorithm can confidently distinguish H_1 and H_2.

Results: Big picture [KSKSR, in submission]



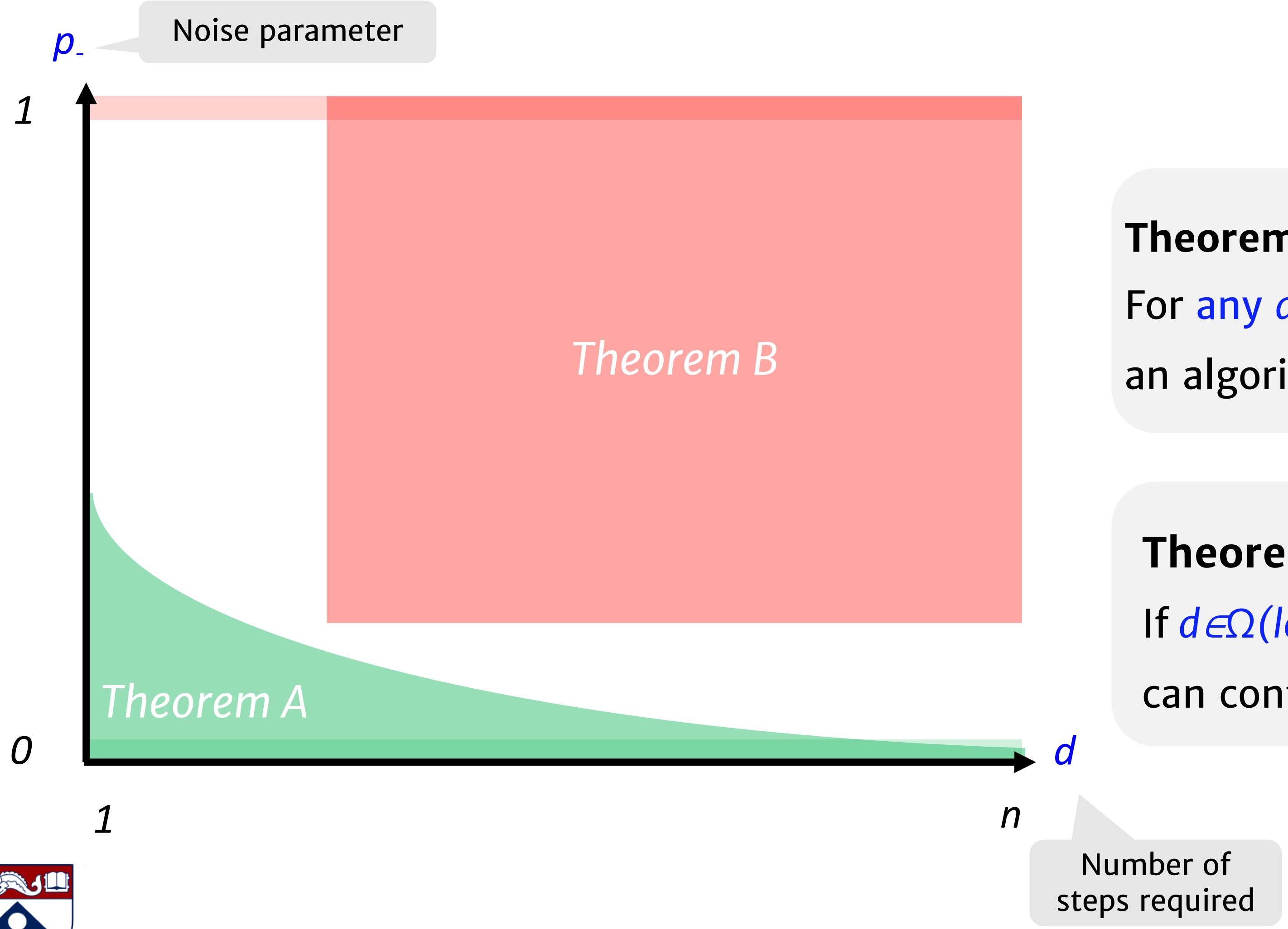
Theorem A (informal)

For any d there is a choice of parameters such that an algorithm can confidently distinguish ***H1*** and ***H2***.

Theorem B (informal)

If $d \in \Omega(\log n)$ and “sufficient” noise, no algorithm can confidently distinguish ***H1*** and ***H2***.

Results: Big picture [KSKSR, in submission]



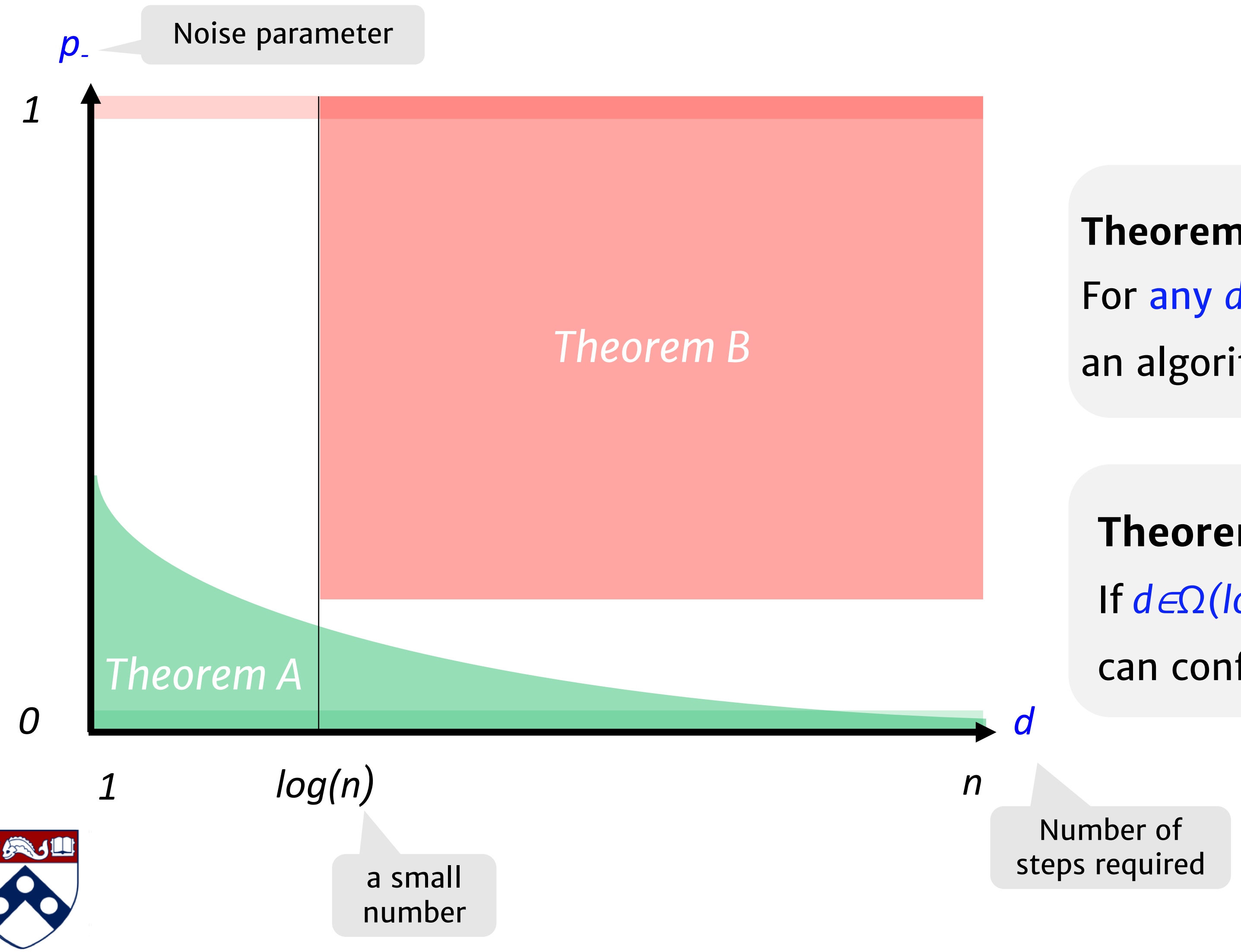
Theorem A (informal)

For any d there is a choice of parameters such that an algorithm can confidently distinguish ***H1*** and ***H2***.

Theorem B (informal)

If $d \in \Omega(\log n)$ and "sufficient" noise, no algorithm can confidently distinguish ***H1*** and ***H2***.

Results: Big picture [KSKSR, in submission]



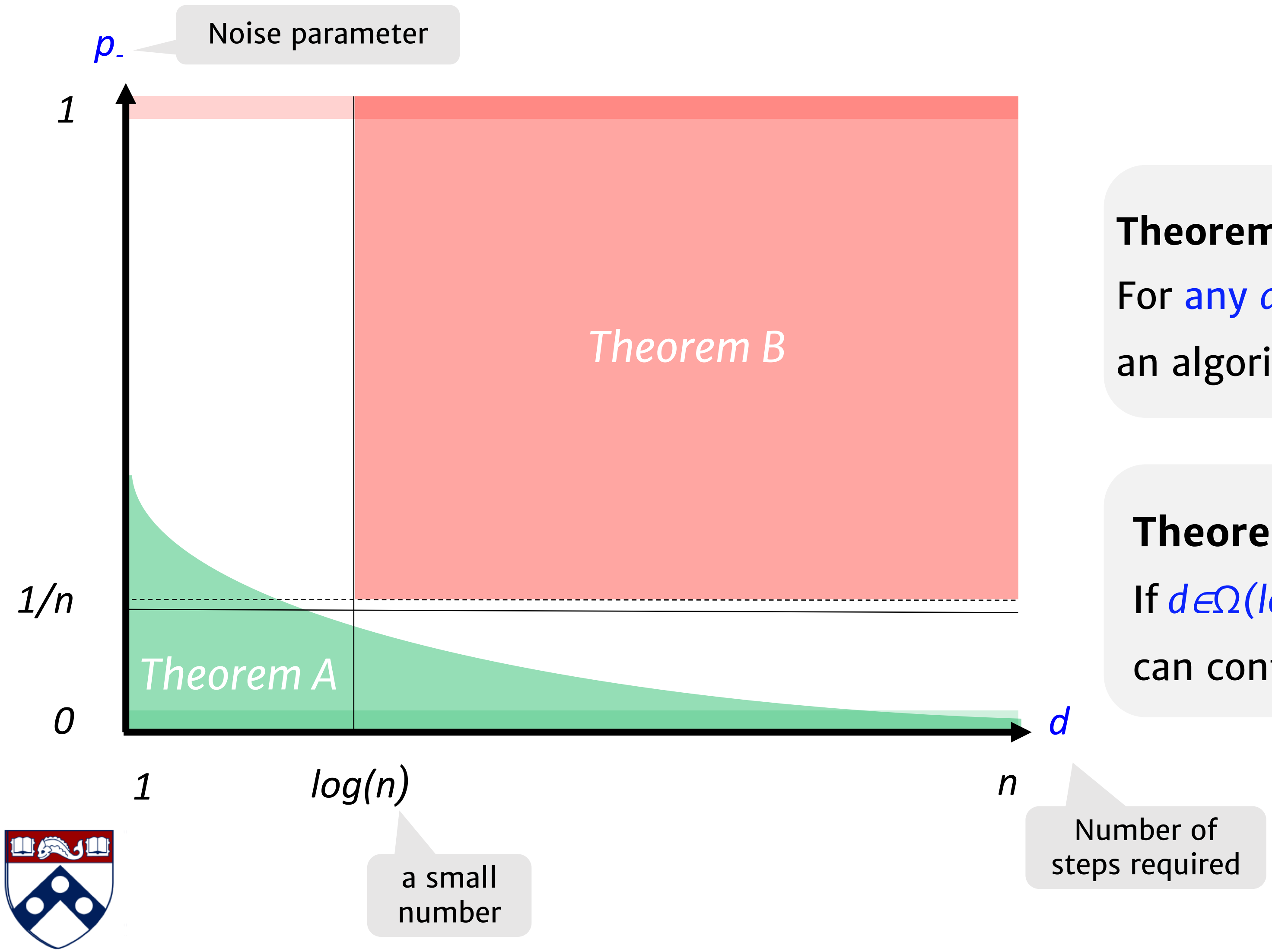
Theorem A (informal)

For any d there is a choice of parameters such that an algorithm can confidently distinguish H_1 and H_2.

Theorem B (informal)

If $d \in \Omega(\log n)$ and "sufficient" noise, no algorithm can confidently distinguish H_1 and H_2.

Results: Big picture [KSKSR, in submission]



Theorem A (informal)

For any d there is a choice of parameters such that an algorithm can confidently distinguish ***H1*** and ***H2***.

Theorem B (informal)

If $d \in \Omega(\log n)$ and "sufficient" noise, <u>no algorithm</u> can confidently distinguish ***H1*** and ***H2***.

Practical lessons

- Pursuing "very long" multi-hop reasoning is unlikely to result in general results.
- Corollary: one has to focus on richer representations (i.e., dealing with **ambiguity** and **variability)** such that it leads to few number of hops needed.

Summary of this section

- A framework for studying "reasoning", in the context of language problems.
- Multi-hop reasoning:
  - \+ There are non-trivial problems where successful reasoning is reliable.
  - \- Reasoning with "large"-many hops likely to fail, even with small amount of noise.
- Implications for practice
  - **Hypothesis:** invest in representations that lead to few hops reasonings.

Summary of the talk

Summary of the talk

- NLU; potentials for significant impacts in the coming years.
- Answering questions: a natural evaluation protocol.
  - Many challenges along the way to this goal: ambiguity, variability, etc.
- Approaches:
  - System design: systems that abstracting over text and reasoning with it.
  - Evaluation: effective benchmarks to measure and incentivize the community.
  - Formalism: to study a class of reasoning algorithms in the context of language.

Summary of the talk

- NLU; potentials for significant impacts in the coming years.
- Answering questions: a natural evaluation protocol.
  - Many challenges along the way to this goal: ambiguity, variability, etc.
- Approaches:
  - System design: systems that abstracting over text and reasoning with it.
  - Evaluation: effective benchmarks to measure and incentivize the community.
  - Formalism: to study a class of reasoning algorithms in the context of language.

Summary of the talk

- NLU; potentials for significant impacts in the coming years.
- Answering questions: a natural evaluation protocol.
  - Many challenges along the way to this goal: ambiguity, variability, etc.
- Approaches:
  - System design: systems that abstracting over text and reasoning with it.
  - Evaluation: effective benchmarks to measure and incentivize the community.
  - Formalism: to study a class of reasoning algorithms in the context of language.

Summary of the talk

- NLU; potentials for significant impacts in the coming years.
- Answering questions: a natural evaluation protocol.
  - Many challenges along the way to this goal: ambiguity, variability, etc.
- Approaches:
  - System design: systems that abstracting over text and reasoning with it.
  - Evaluation: effective benchmarks to measure and incentivize the community.
  - Formalism: to study a class of reasoning algorithms in the context of language.

Summary of the talk

- NLU; potentials for significant impacts in the coming years.
- Answering questions: a natural evaluation protocol.
    - Many challenges along the way to this goal: ambiguity, variability, etc.
- Approaches:
    - System design: systems that abstracting over text and reasoning with it.
    - Evaluation: effective benchmarks to measure and incentivize the community.
    - Formalism: to study a class of reasoning algorithms in the context of language.

Summary of the talk

- NLU; potentials for significant impacts in the coming years.
- Answering questions: a natural evaluation protocol.
  - Many challenges along the way to this goal: ambiguity, variability, etc.
- Approaches:
  - System design: systems that abstracting over text and reasoning with it.
  - Evaluation: effective benchmarks to measure and incentivize the community.
  - Formalism: to study a class of reasoning algorithms in the context of language.

Summary of the talk

- NLU; potentials for significant impacts in the coming years.
- Answering questions: a natural evaluation protocol.
  - Many challenges along the way to this goal: ambiguity, variability, etc.
- Approaches:
  - System design: systems that abstracting over text and reasoning with it.
  - Evaluation: effective benchmarks to measure and incentivize the community.
  - Formalism: to study a class of reasoning algorithms in the context of language.

Thesis Publication

- KSKSR. *On the Capabilities and Limitations of Reasoning for Natural Language Understanding*, in submission.
- ZKNR. *A Question Answering Benchmark for Temporal Common-sense*, under review.
- KCRUR. *Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences*, NAACL, 2018.
- KKSR. *Question Answering as Global Reasoning over Semantic Abstractions*, AAAI, 2018.
- KKSR. *Learning What is Essential in Questions*, CoNLL, 2017.
- KKSR. *Question Answering via Integer Programming over Semi-Structured Knowledge*, IJCAI, 2016.

Other Publications

- NLP:
  - C**K**WCR. *Seeing Things from a Different Angle: Discovering Diverse Perspectives about Claims*, under review.
  - Z**K**CR. *Zero-Shot Open Entity Typing as Type-Compatible Grounding*, EMNLP, 2018.
  - CEKSTT**K**. *Combining Retrieval, Statistics, and Inference to Answer Elementary Science Questions*, AAAI, 2016.
  - F**K**PWR. *Illinois-Profiler: Knowledge Schemas at Scale*, Cognitum, 2015.
  - P**K**R. *Solving Hard Co-reference Problems*, NAACL, 2015.
- NLP software/tools:
  - **K** et al. *CogCompNLP: Your Swiss Army Knife for NLP*, LREC, 2018.
  - SC**K**KSVBWR. *EDISON: Feature Extraction for NLP, Simplified*, LREC, 2016.
- ML/Optimization/etc:
  - KS**K**CSSR. *Relational Learning and Feature Extraction by Querying over Heterogeneous Information Networks*, StartAI, 2018.
  - K**K**CMSR. *Better call Saul: Flexible Programming for Learning and Inference in NLP*, COLING, 2016.
  - Q**K**. *Online Learning with Adversarial Delays*, NourIPS, 2015.
  - **K**NJF. Joint Demosaicing and Denoising via Learned Non-parametric Random Fields, TIP, 2014.


Snigdha Chaturvedi
(UCSC)


Tushar Khot
(AI2)


Dan Roth
(UPenn)

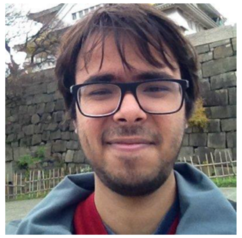

Ashish Sabharwal
(AI2)


Erfan Sadeqi Azer
(Indiana U)


Oren Etzioni
(AI2)

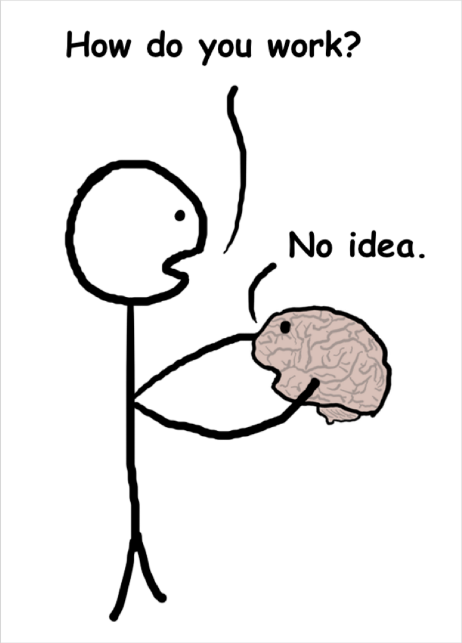

Peter Clark
(AI2)


Michael Roth
(Saarland U)


Shyam Upadhyay
(Upenn → Google)


Ben Zhou
(UIUC → UPenn)

- That’s it folks



Questions?