Event-centric Context Modeling:

The Case of Story Comprehension and Story Generation

Nasrin Mostafazadeh

elemental cognition

June 2018

State of Artificial Intelligence, ~12 years ago

RoboCup Competitions

Classic NLU Example

- The monkey ate the banana because it was hungry.
  - Question: What is it? Monkey or the banana?
  - Correct answer: Monkey

Deemed very challenging for AI systems at the time

Robotics vs NLU

Boston Dynamics' Most Recent Robot

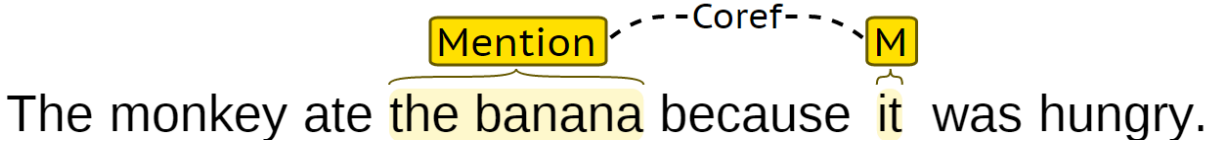
(Feb 2018)

Stanford CoreNLP Coreference Resolver

(June 2018)

Classic Example:

- The monkey ate the banana because it was hungry.
  - What is it? Monkey or the banana?

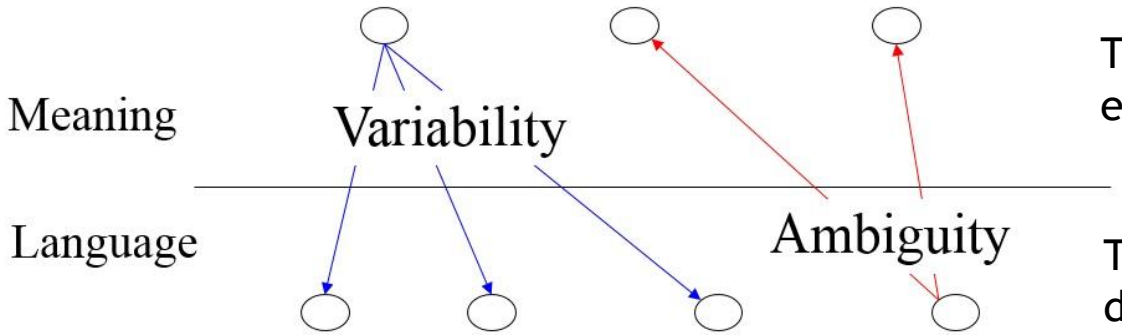


Slide credit: Omid Bakhshandeh

Why is NLU Hard?

- The Classic Dual Problem of Language Ambiguity and Meaning Variability

Vast amount of Knowledge



The same meaning can be expressed in many ways

The same expression can mean different things and be ambiguous.

Language Understanding in Eventful Contexts

Human-level Understanding in Context

Context: At the grocery store

- **Customer**: Black beans?
- **Clerk**: Aisle 3.

Context: Serving food

- **Woman**: Black beans?
- **Man**: Yeah, I love it.

Example Credit: Philip Cohen and James Allen

Context: Serving food

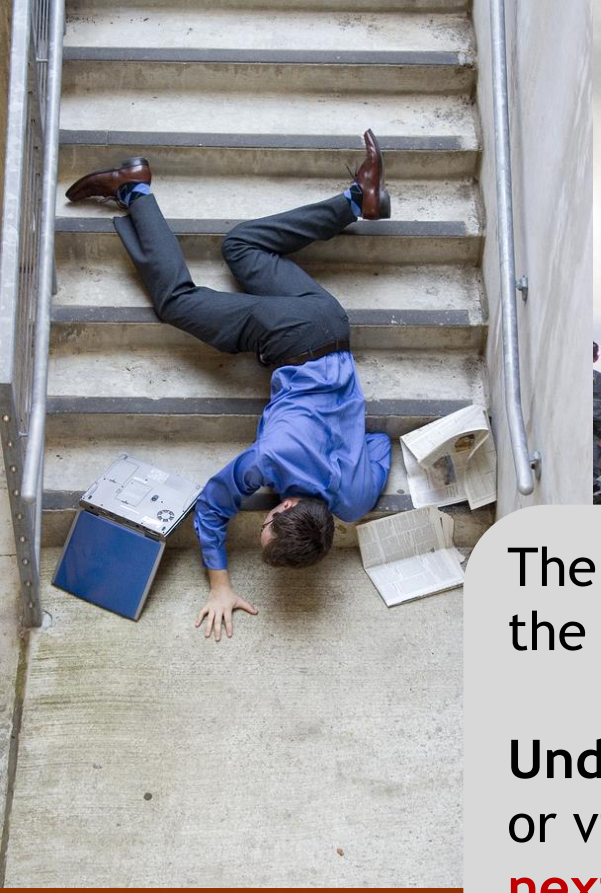
- **Man**: Black beans?
- **Woman**: Oh, you don’t like it?

Fully understanding the underlying linguistic context (no matter how simple) requires the integration of an agent's perception (speech, text, vision, etc.) with its:

- World model
  - Different parties' beliefs and desires
  - The dynamics of events
- Intention Recognition
- Planning
- ...

Event-centric Context Modeling: The Case of Story Comprehension and Story Generation

With a focus on commonsense reasoning



The changes of the world are **caused** by the **effects** of **events**.

Understanding events through language or vision, and **predicting what happens next**, is one of the **most demanding areas in AI**.

Storytelling has been one of the oldest ambitions of AI, and one of the most extremely challenging tasks!

Story Understanding

Story Generation

Narrative Structure Learning

Narrative Intelligence

Episodic Knowledge Acquisition

Script Learning

Storytelling

Schank and Abelson, 1977; Dyer, 1983; Charniak 1972; Turner, 1994; Schubert and Hwang, 2000, …

We should build AI systems that can communicate through stories!

- Narrative is a major cognitive tool humans use for holding meaningful communications (Dahlstrom, 2014; AC, 2002), serving a variety of purposes.
- Evidence suggests that audiences better understand and remember narratives, as opposed to expository writing.

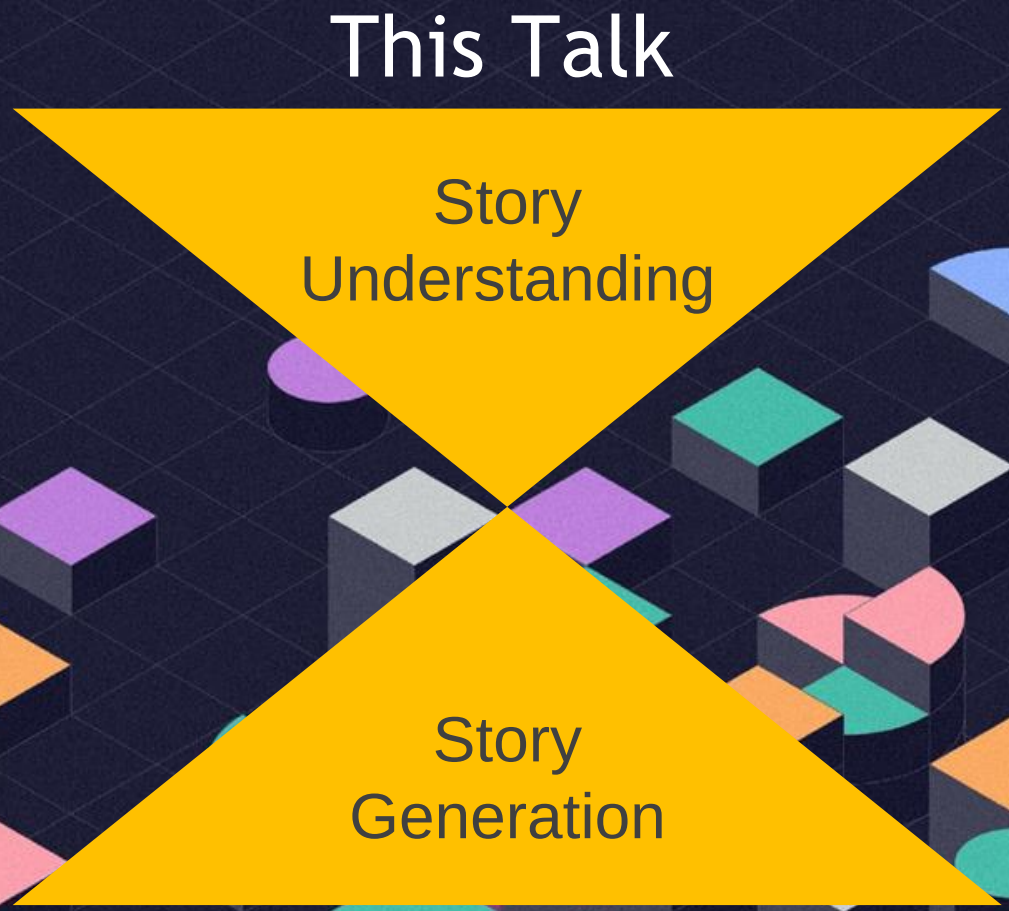
This Talk:

Story of my research on

- Story Comprehension
- Story Generation

The Story
about Stories!
#cheesynaacl

- Both extremely challenging tasks in AI.
- **Biggest challenge**: commonsense knowledge for the interpretation of narratives.



Modeling Textual Narrative Context

Goal: Building a system that can **comprehend stories**

What is my definition of “story”?

- “A narrative or story is anything which is told in the form of a logically linked set of events”
  - At this point we are not concerned with how entertaining or dramatic the stories are!

Resources

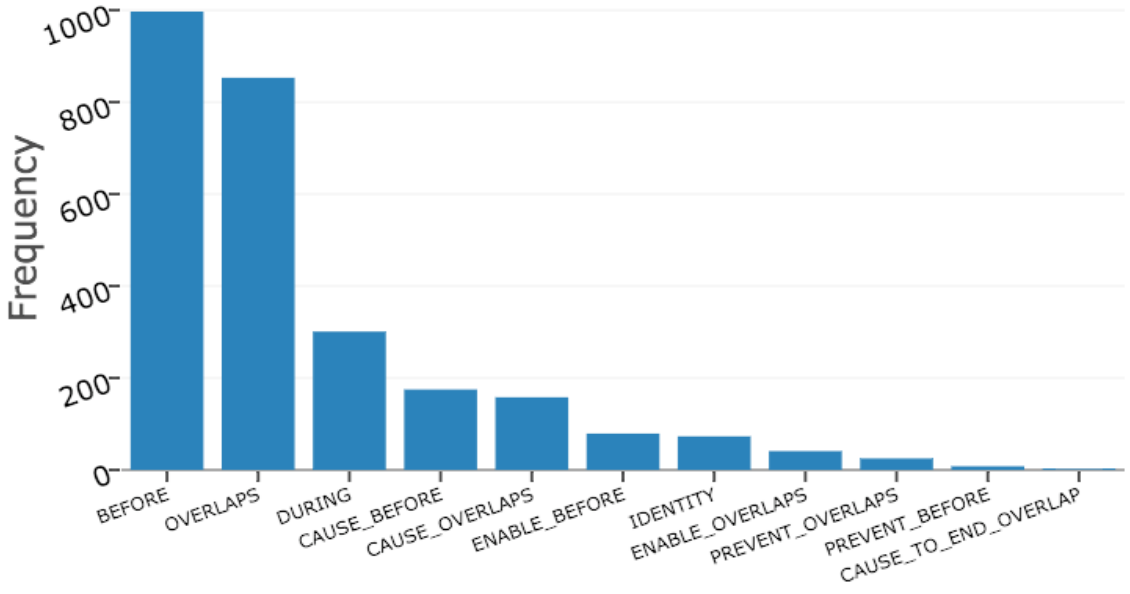
ROCStories

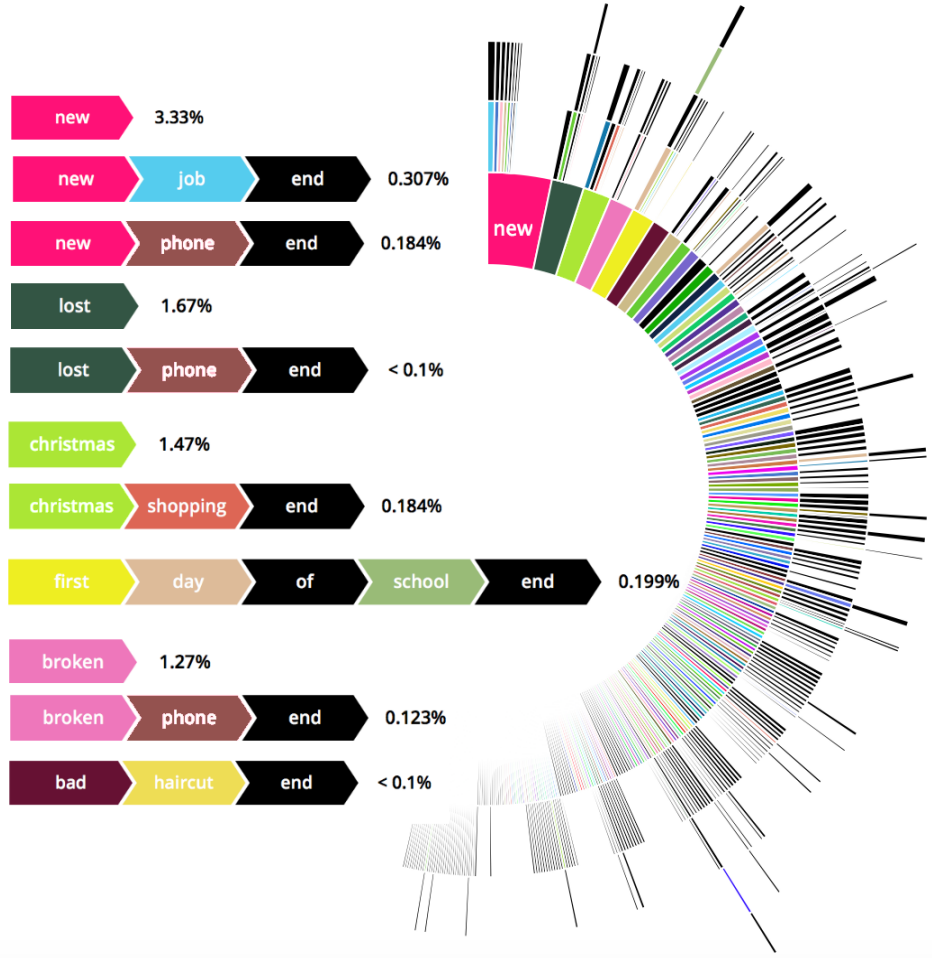
Mostafazadeh et al., NAACL 2016

Where to Start Learning Stories/ Narrative Structures From?

ROCStories: Short Commonsense Stories

- A collection of 100K high quality short realistic five-sentence stories with their titles authored by hundreds of crowd workers.





An Example ROC Story
Title: “A Friendly Game”

Bill thought he was a great basketball player.
He challenged Sam to a friendly game.
He agreed.
Sam started to practice really hard.
Eventually, Sam beat Bill by 40 points.

X challenges Y —enable→ Y agrees to play —before→ Y practices hard —enable→ Y beats X

Causal and Temporal Relation Scheme (CaTeRS), Mostafazadeh et al., Event Workshop at NAACL 2016

How to do automatic evaluation for story understanding?

Research had been hindered by the lack of a proper evaluation framework!

The Idea: Story Cloze Test (SCT)

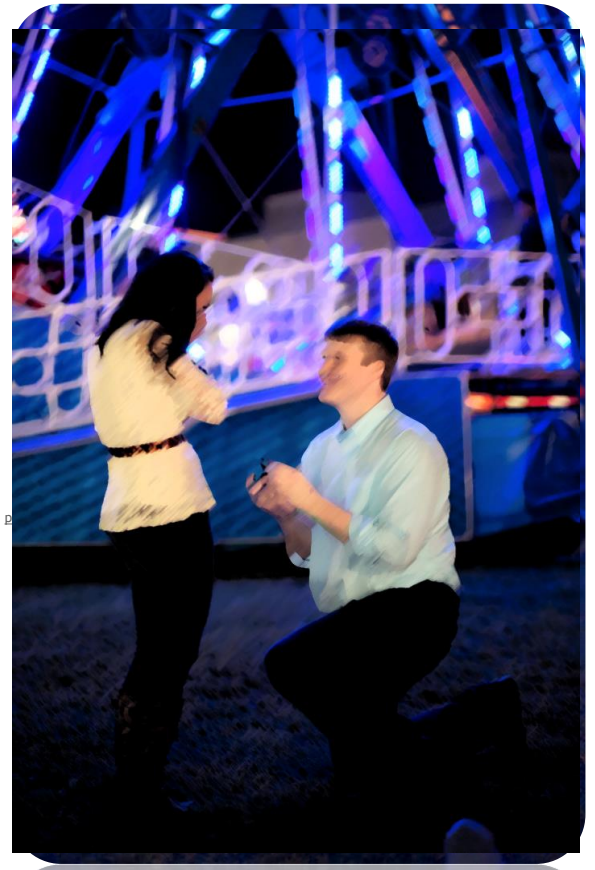
- **Goal**: Design a new evaluation schema for story understanding and narrative structure learning.
- **The Story Cloze Test**: Given a context of four sentences and two alternative endings to the story, choose the correct ending.
  - Scale the evaluation by crowdsourcing the dataset

Predicting what happens next

An Example Story Cloze Test

- **Context:** Tom and Sheryl have been together for two years. One day, they went to a carnival together. He won her several stuffed bears, and bought her funnel cakes. When they reached the Ferris wheel, he got down on one knee.
- **Right Ending:**
  - Tom asked Sheryl to marry him.
- **Wrong Ending:**
  - He took out his boots and sat down.

Goal: Enforce deep language understanding



We collected 3,744 **doubly human-verified** Story Cloze Test instances

Story Cloze Test: The benchmark for narrative understanding

- Human performs 100%
- A challenging task with a wide enough gap (42%) from the state-of-the-art and human performance, so plenty of room for research!

- Various use-cases
  - Training models which understand or tell stories
  - Training generic language models
  - Evaluating children's intellectual disabilities!
  - Developing theories of what makes a sequence a story.
  - …
- All resources related to the ROCStories project http://cs.rochester.edu/nlp/rocstories/

Story Cloze Models

Story Cloze Baselines

$$Accuracy = \frac{\#\ Correct}{\#\ All\ test\ cases}$$

	Constant-choose-first	Frequency	N-gram-overlap	GenSim	Sentiment-Full	Sentiment-Last	Skip-thoughts	Narrative-Chains-AP	Narrative-Chains-Stories	DSSM	Human
Validation Set	0.514	0.506	0.477	0.545	0.489	0.514	0.536	0.472	0.510	0.614	1.0
Test Set	0.513	0.520	0.494	0.539	0.492	0.522	0.552	0.478	0.494	**0.595**	1.0

The 1st Story Cloze Shared Task

Mostafazadeh et al., LSDSem @ EACL 2017

1.00 → Human Performance

CodaLab

Results

#	User	PercentageScore ▲
1	msap	0.752004 (1)
2	cogcomp	0.743987 (2)
3	tbmihaylov	0.724212 (3)
4	ukp	0.716729 (4)
5	Niko	0.700160 (5)
6	roemmele	0.671833 (6)
7	mflor	0.620524 (7)
8	Pranav_Goel	0.604490 (8)
9	ROCNLP	0.595938 (9)
10	lizhongyang	0.585249 (10)
11	sjtuadapt	0.585249 (10)

The winner

Baseline

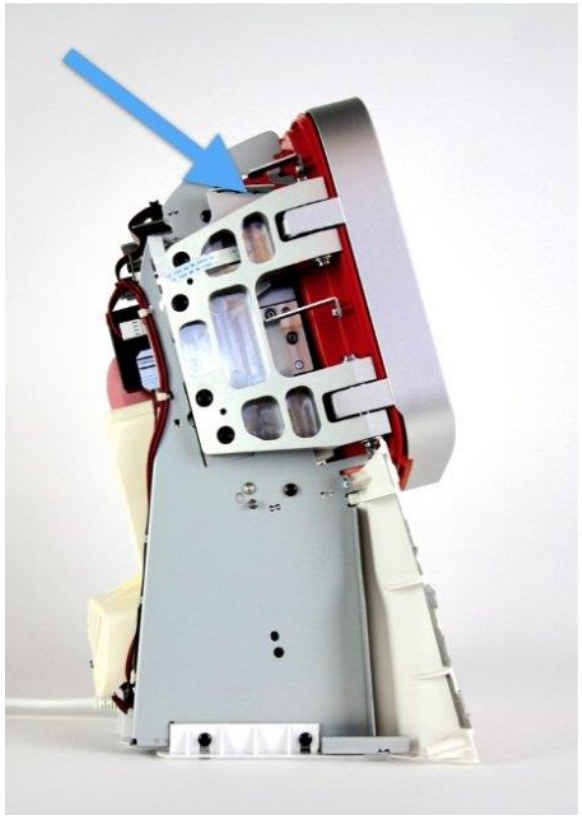
elemental cognition

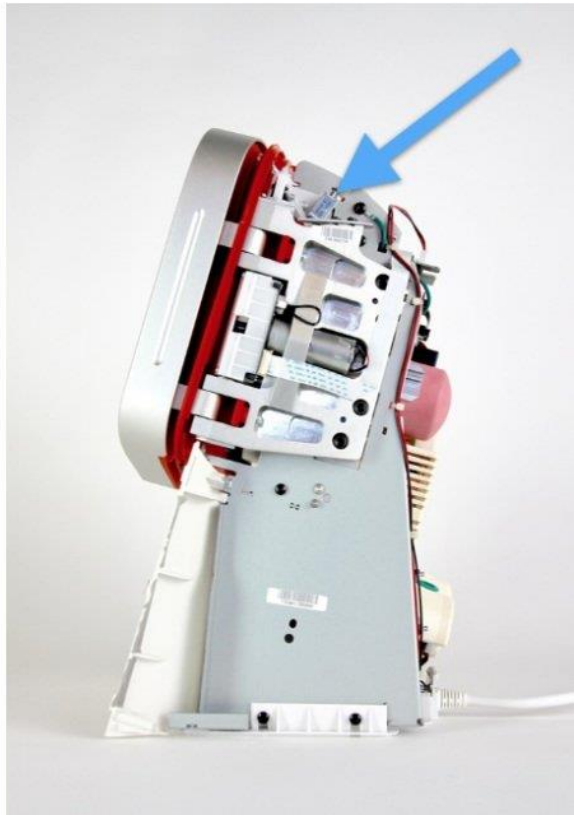
The Shared Task: Notable Trends

Hey, Juicero!

elemental cognition

Beautiful Engineering

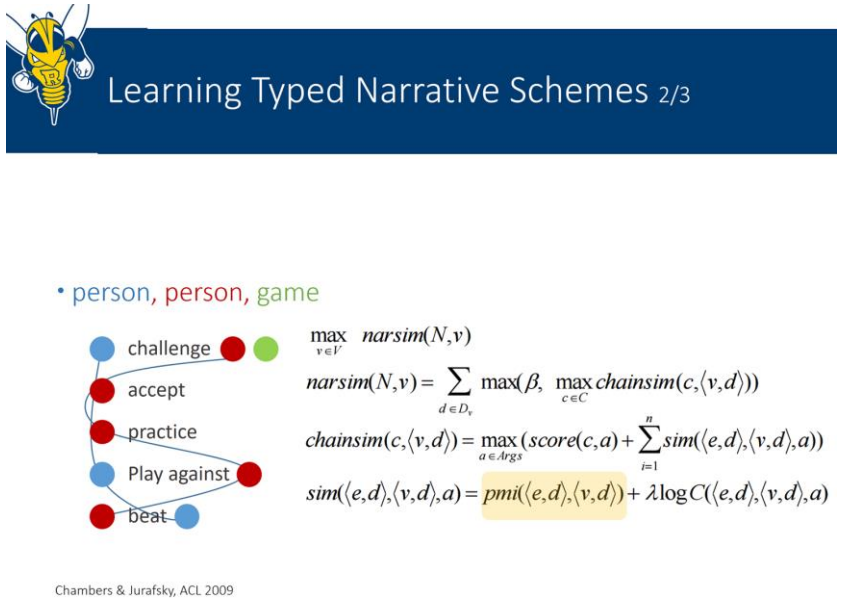




elemental cognition

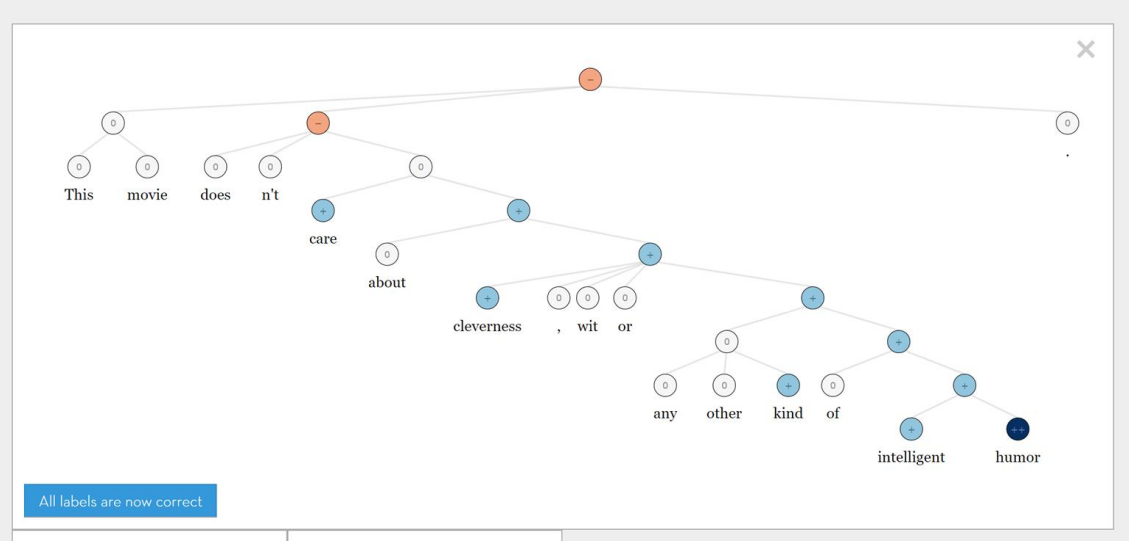
& Our Obsession with Complexity …

Model Complexity … 1/2



- Romelle et al. (2017) computed basic PMI score for all the word pairs of context: achieve 59.9 vs our 49.4

Our Love for Model Complexity … 2/2



We used “Recursive Neural Network” sentiment analyzer trained on ~12,000 sentences

- Goel & Singh (2017) use VADER (a rule-based sentiment analyzer) for sentiment-match and achieve 58.2 vs ours 49.2

- - Although we were very careful with our task design, data collection process, and establishing various baselines, it was shown (Schwartz et al., 2017; Cai et al., 2017) that our writing task had imposed its own biases on our dataset
  - * Linguistic “stylistic” features of only the ending sentences (correlated with deceptive text features)
    - - Sentence length
    - - Use of pronouns and adjectives
    - - Word & character n-gram.

- - From NLI, to VQA, Story Cloze Test, our narrow benchmarks inevitably have data creation artifacts & hence yield biased models. If we're lucky, we find some of the artifacts,while many stay hidden.

Very **crucial** to discover **hidden data biases** in various AI tasks.

Tackling The Stylistic Biases in the Story Cloze Test
(Sharma et al., ACL 2018)

- - Our ACL2018 paper tackles some of the biases in the Story Cloze Test. Stay tuned for the new data release: http://cs.rochester.edu/nlp/rocstories/
- - Despite all the efforts, we emphasize that “no curated dataset is without its inherent hidden biases” and suggest building benchmarks iteratively & report performance on all.
  - * Remember that even the SOTA on the original Story Cloze Test V1.0 is far from human performance!

The SOTA (CogComp) performance goes from 77% on Story Cloze Test v1.0 to 60.8% on v1.5.

Classification makes for a great benchmark, but is prone to hacking!

How to Enforce Deep Language Understanding?

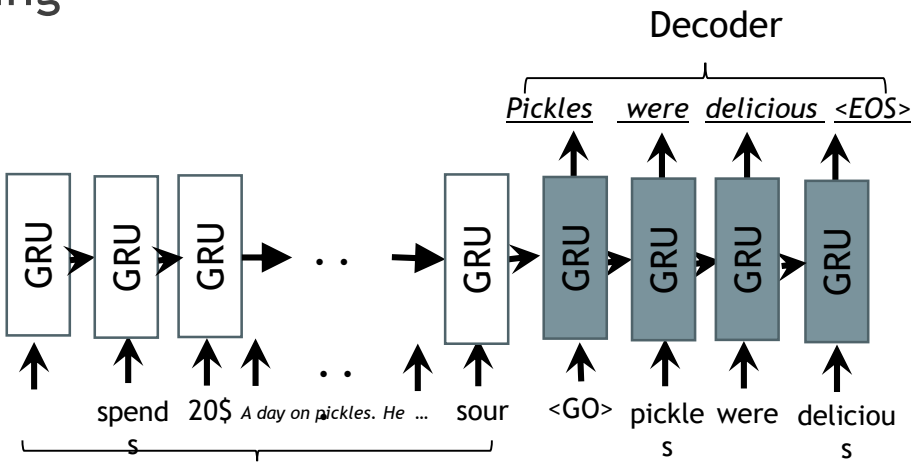
1. Story Generation

Goal: Given a narrative context, the system and human should collaboratively tell a logically-sound story

The AI system should generate the ending as opposed to choosing it.

The Case Study: Sequence2Sequence Generation
(Sutskever et. al, 2014)

- Encoder-Decoder RNN Architecture with Attention
- Beam-search decoding



Simply fails when feeding in all the four sentences as the context for generating the ending

What if we set the context to be only the 4th sentence?

- 4th Sentence: She wrote another one through self-publishing avenues.
- System: And they laughed !

- 4th Sentence: She got it painted a bright pink color.
- System: Kelly is very happy .

- 4th Sentence: He did extra credit assignments.
- System: Everyone loved it .

- 4th Sentence: At midnight, he looked back and saw Anita dancing with another guy.
- System: He was thrilled when he was the man he wanted to be !

It is easy to generate something that kind of makes sense … when you **discard the bigger context**!

elemental cognition

The importance of Modeling (non-trivial) Context

As humans,
we have the amazing capability of **imagining a scenario** for **validating** even a **random output** whenever there is no/short established prior context!

Context Representation: Simple Generalizations

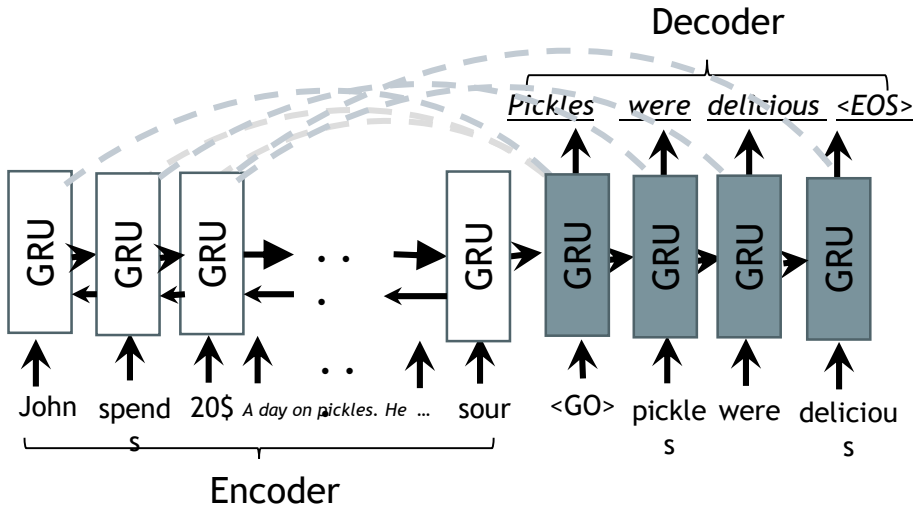
- Retaining only the main verbs in the context

Input 4-sentence story (only highlighted words)**:**

John spends $20 a day on pickles. He decides to make his own to save money. He puts the pickles in brine. John waits 2 weeks for his pickles to get sour.

Output fifth-sentence:
Pickles were delicious <EOS>

<EOS>= end of sentence marker
- We extract the **main events and their semantic roles** using TRIPS STEP parser.



- Bi-Directional Encoder-Decoder RNN Architecture with Attention
- 2-layers, with 512 units per layer
- Beam-search decoding, with beam-width = 25
- Reranking using PRO algorithm

elemental cognition

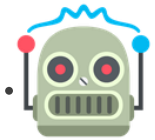
Example Generation

- **Context:** John spends $20 a day on pickles. He decides to make his own to save money. He puts the pickles in brine. John waits 2 weeks for his pickles to get sour.

- 1: Now he is so happy that he has money. 
- 2: John opens the jar to find perfect pickles. 

Example Generation

- **Context:** Kara lost her dog yesterday. She freaked out. Frantic, Kara looked all over town for it. Finally she found the dog in her neighbor's yard.

- 1: She was relieved to have it back home. 
- 2: She asked the neighbors not to put out treats. 

But Often …

- Context: Ivy wanted a cell phone like all her friends. She saved up her money from delivering papers. Soon she had enough for a phone. She went shopping and selected the perfect phone.
- System: She was so happy she was gotten her food !

- Context: Nicole wanted to be an Aunt more than anything. The trouble was that she didn't have any siblings. Today Nicole got a note from her cousin, Xiao. Xiao and her husband were expecting a baby boy.
- System: He was glad he did not .

Context Representation: Better Generalization

Preprocessing:
- NER
- Coreference Resolution
- Abstraction using Ontology Type

- John spends $20 a day on pickles. He decides to make his own to save money. He puts the pickles in brine. John waits 2 weeks for his pickles to get sour.

- PERSON1 ONT::commerce-pay $20 a day on ONT::condiment. PERSON1 ONT::decide to ONT::create PERSON1* to ONT::save-cost ONT::money. PERSON1 puts the ONT::condiment in ONT::brine. PERSON1 ONT::waits DURATION1 for PERSON1* ONT:condiment to ONT:become ONT:sour.

Collaborative Turn-by-Turn Generation

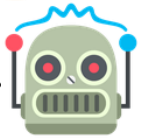
- **PERSON1** ONT::commerce-pay $20 a day on ONT::condiment.

- **PERSON1** decided to go to the store.

- **PERSON1** ONT::purchase more ONT:condiment.

- **PERSON1** was very happy.

Generate the Ending

- **PERSON1** ONT::commerce-pay $20 a day on ONT::condiment. **PERSON1** ONT::decide to ONT::create **PERSON1*** to ONT::save-cost ONT::money. **PERSON1** puts the ONT::condiment in ONT::brine. **PERSON1** ONT::waits **DURATION1** for **PERSON1*** ONT:condiment to ONT:become ONT:sour. 
- **PERSON1** was very proud. 

Language Generation: Where are we standing?

- RNNLMs are performing great on generating grammatical outputs
  - Maintaining local coherency
- Logically-sound generation is still very challenging
  - Generation given a trivial context (a topic, or a title) is easier than generating a logically-sound output given a non-trivial long context
- Generating Shakespeare-like text or poetry is less challenging
    - Often, irrelevant content can be deemed “creative” by human!

What is still very hard?
“to generate a **contentful** sequence of **logically related** sentences.”

A crucial requisite for story comprehension and generation: Better Narrative Context Representation

- We need models that learn to ‘generalize’ better
  - Any training corpus for a generation task requiring commonsense knowledge will be small, if we don’t work on better ‘abstractions’!
  - We should leverage semantic abstractions for better context representation

Generalization & Abstraction is a bottleneck across different AI tasks!

Mostafazadeh et al., ACL 2016

- **Human**: Did the drivers of this accident live through it?
- **GRNN**: How did the car crash?
- **KNN**: Was anybody hurt in this accident?
- **Caption Bot:** A man standing next to a motorcycle.

Classification makes for a great benchmark, but is prone to hacking!

How to Enforce Deep Language Understanding?

2. Answering ‘Why?’

The truly difficult question that prevents us from getting to human-level AI

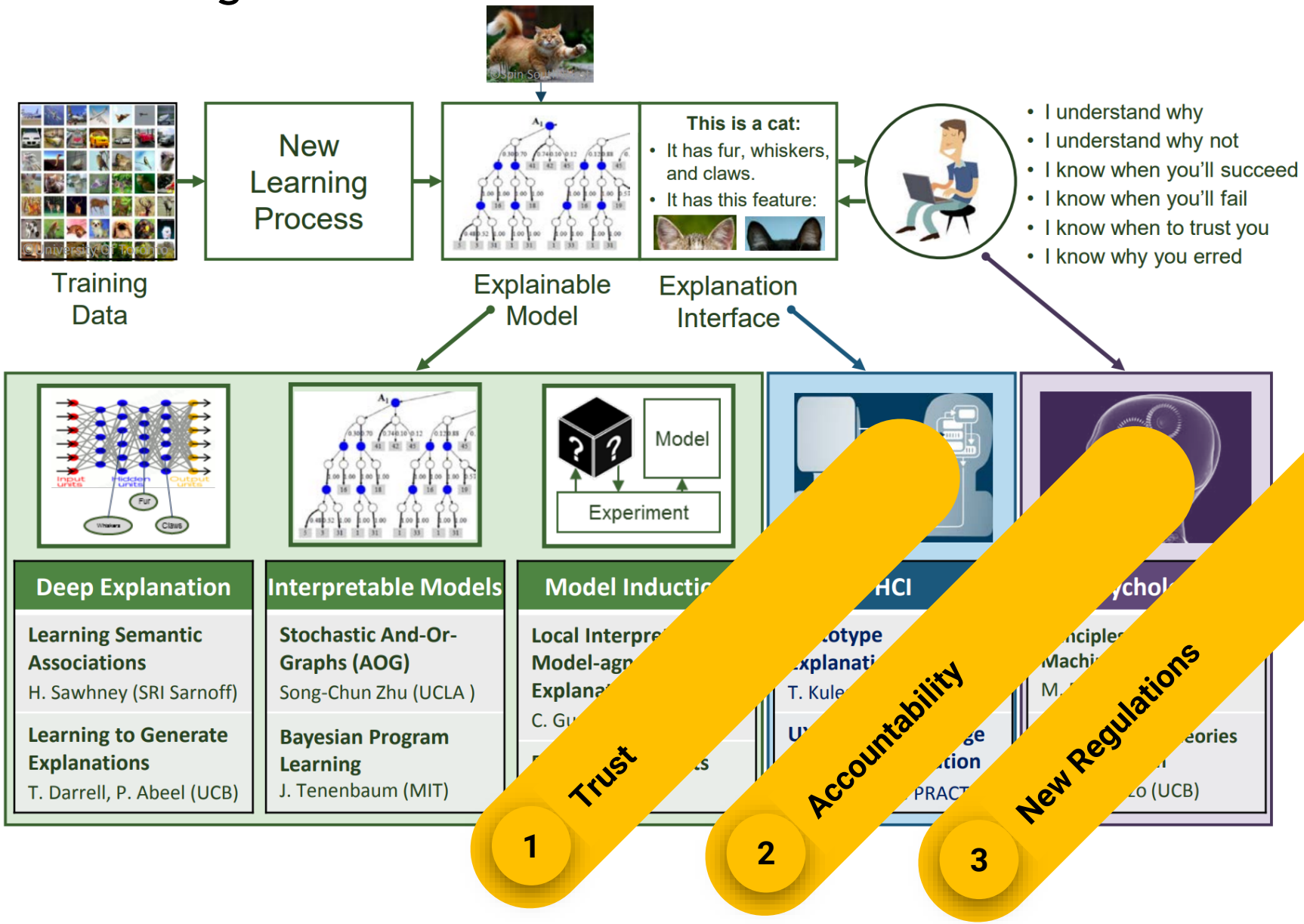
Goal: Given a short narrative, the system comprehend the story and answer a variety of ‘why’ questions about its understanding of the story.

The AI system should **explain** its choice of ending.

Explainable AI is all the rage now

...through the success of the black-box models!

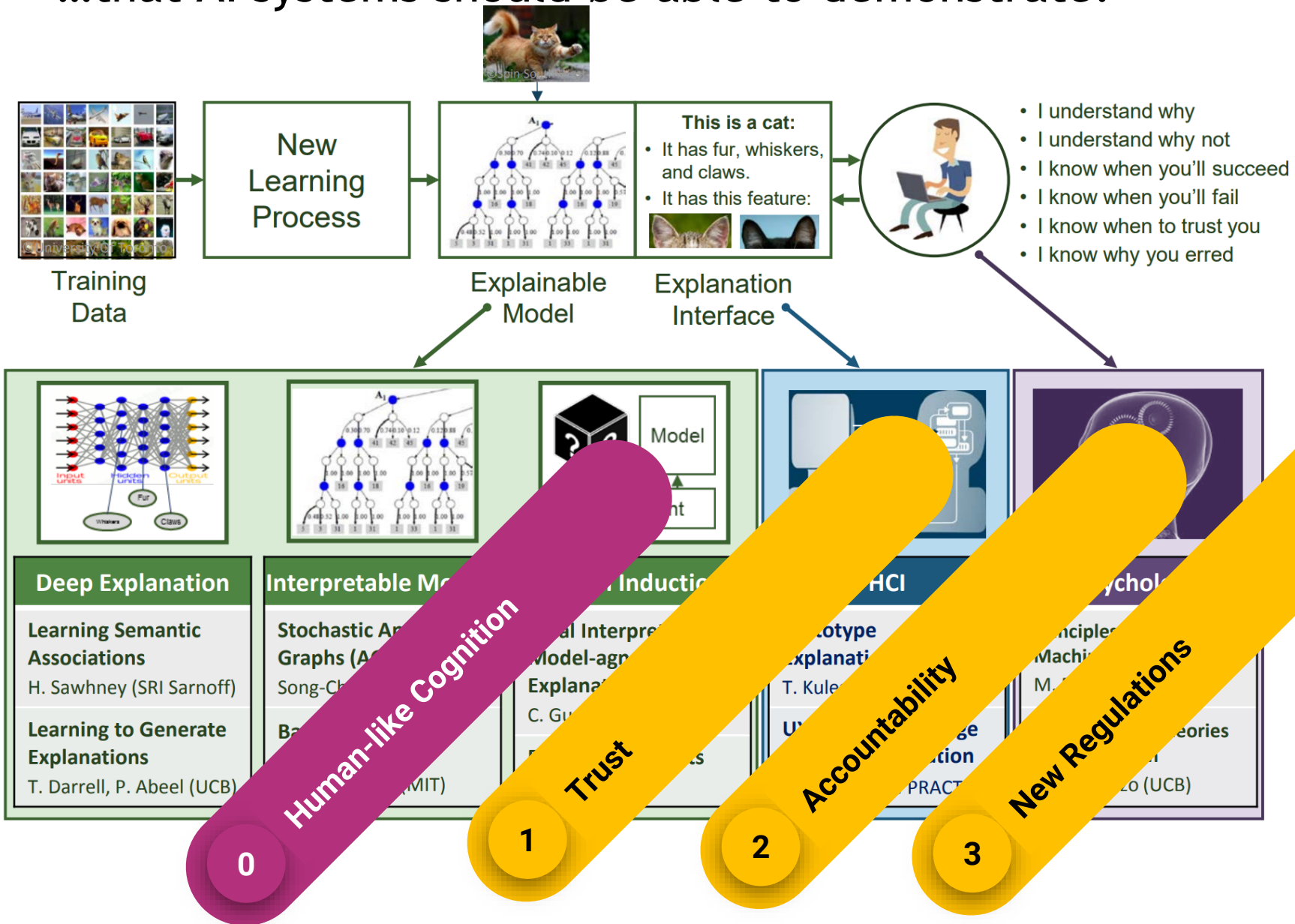
DARPA XAI

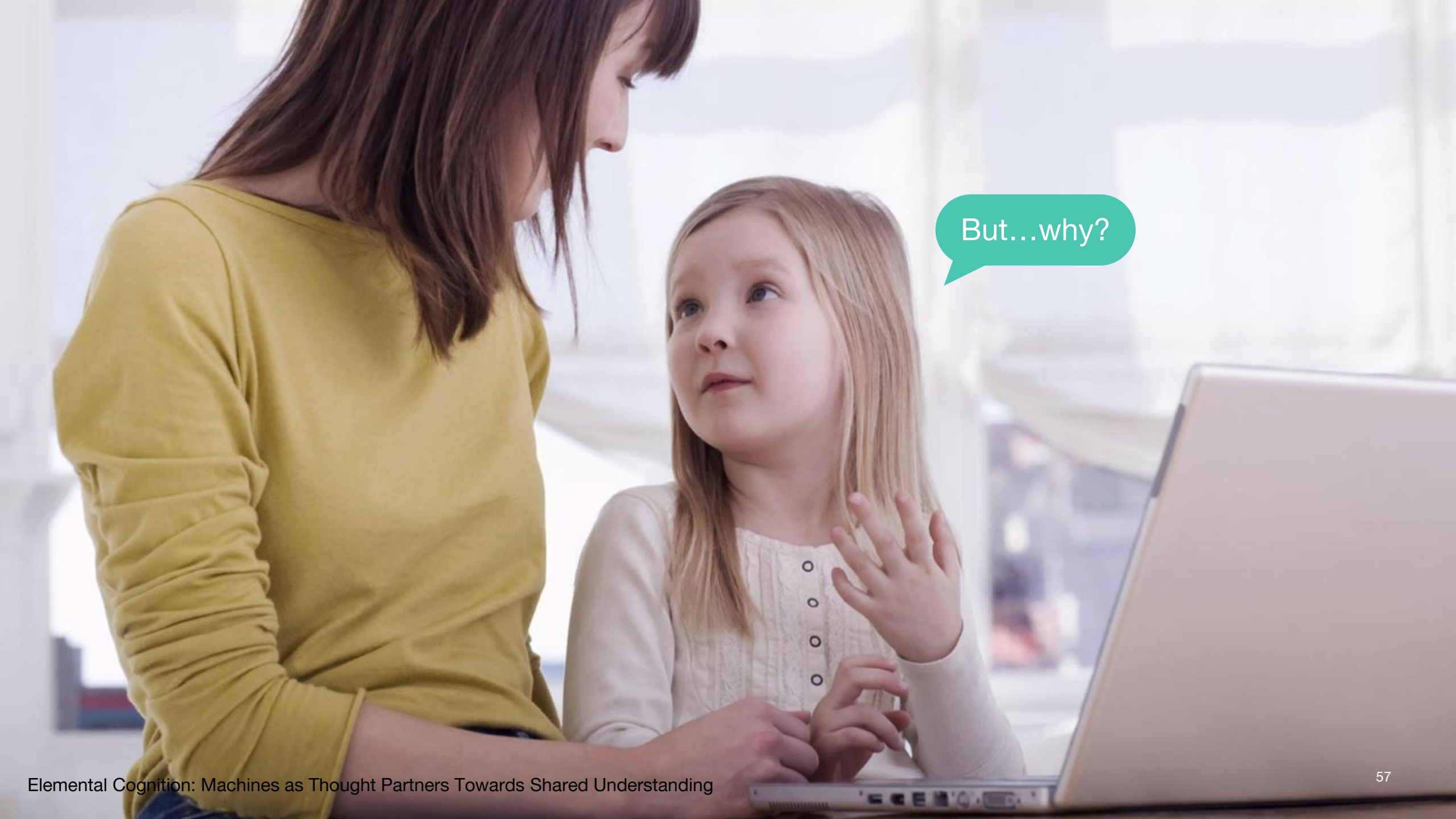


Explanation is a human-like cognitive ability

…that AI systems should be able to demonstrate!

DARPA XAI




But…why?

Lucky Louie S01 E01: https://youtu.be/D4JkD_7IEBM

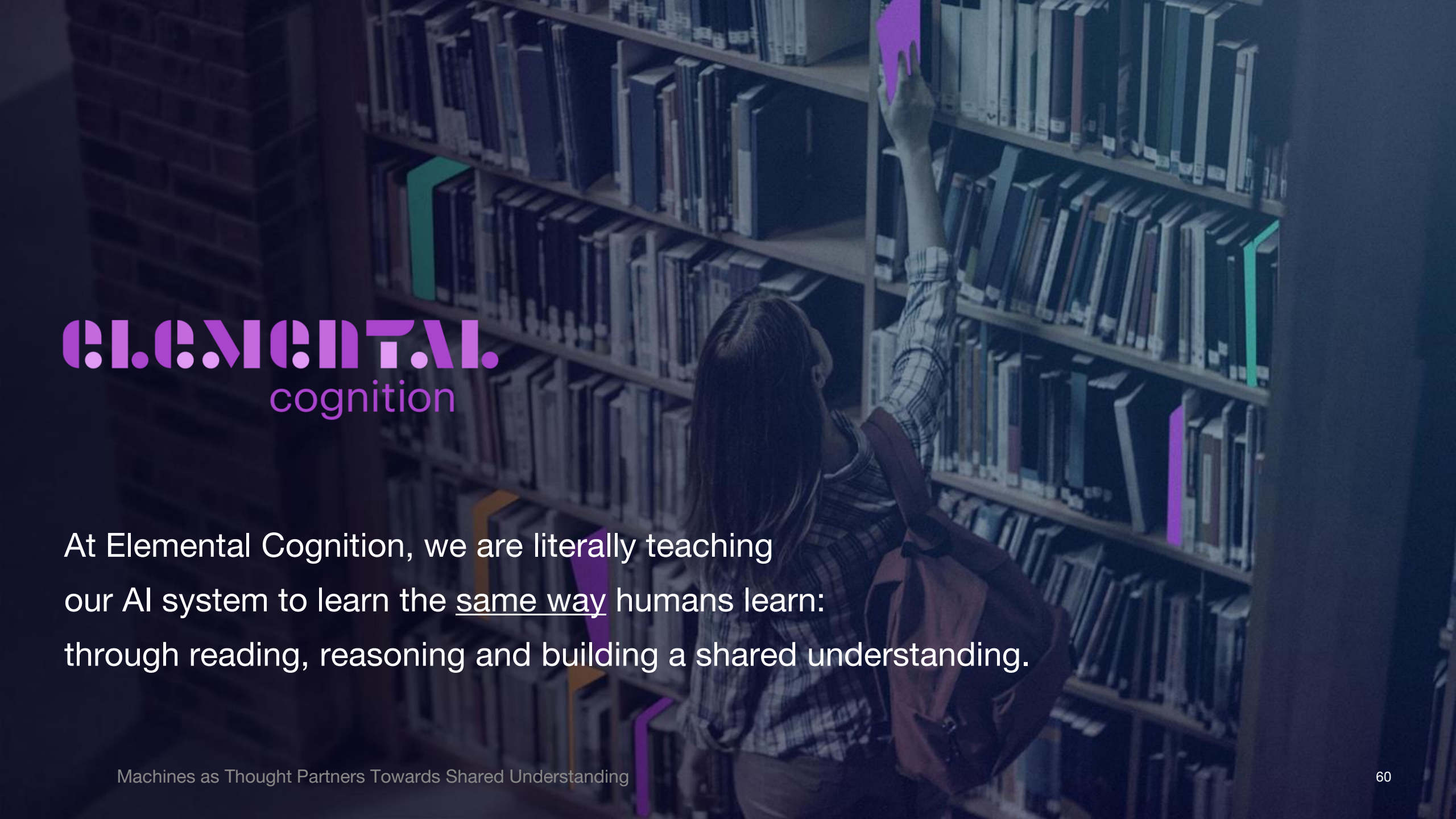
ELEMENTAL
cognition

On the Quest to Answering "Why"

Our Moonshot:
Machines as Thought Partners

Towards a Shared Understanding

Through Machine Reading, Reasoning, and Human-Machine Collaboration



At Elemental Cognition, we are literally teaching
our AI system to learn the same way humans learn:
through reading, reasoning and building a shared understanding.

Our system starts at the beginning…

…and knows it doesn't understand.

The Winning Shot

The soccer game was nearly over. The two teams were tied, one to one.

Alice kicked the ball. Oops! She kicked it the wrong way.

"What a bad kick," Alice thought.

But wait! Her teammate John jumped to stop the ball. The ball went off John and into the goal! Alice's team won the game. Hooray! It was not a bad kick, after all.

How many points did Alice's team have?

Answer: I don't know.
✕ Incorrect

Explanation: I don't know.
✕ Incorrect

1 2

I didn't understand very well.

I got some questions right, but couldn't explain my answers.

I think it would help if I read another story about soccer first.

It looks to learn from the ground up.

How to Play Soccer

Soccer is a team sport. Two teams play. Each team has eleven players. The players try to kick the ball into the other team's goal. If the ball goes into the goal, their team gets a point.

Soccer players have different jobs. The goalkeeper's job is to protect the goal. Defenders try to get the ball away from the other team. Forwards try to kick the ball into the other team's goal. Games usually last ninety minutes. The team with the most points is the winner!

Ok, I found a story about the game of soccer. Take a minute to read this one.

OK, I'm done.

The system identifies concepts it doesn't know about and draws on large text corpora to fill in the gaps.

It says, "Soccer is a team sport," but I don't understand what a "team sport" is. I read some other stories and found the following sentences. Which is most related to "team sport?"

A. A coach focuses on hockey.

B. Two teams compete with each other in a match.

C. A player grows up to play soccer.

Submit

It then looks to learn from the ground up.

How to Play Soccer

Soccer is a team sport. Two teams play. Each team has eleven players. The players try to kick the ball into the other team's goal. If the ball goes into the goal, their team gets a point.

Soccer players have different jobs. The goalkeeper's job is to protect the goal. Defenders try to get the ball away from the other team. Forwards try to kick the ball into the other team's goal. Games usually last ninety minutes. The team with the most points is the winner!

B championship

C title

Submit

Ok. The sentence said, “Two teams compete with each other in a match.” Which of the following do you think are correct?

A Each team wants to win.

B Each team wants the other team to win.

C Each team wants the other team not to win.

Submit

After some dialog, the system understands what it read about soccer.

How to Play Soccer

Soccer is a team sport. Two teams play. Each team has eleven players. The players try to kick the ball into the other team's goal. If the ball goes into the goal, their team gets a point.

Soccer players have different jobs. The goalkeeper's job is to protect the goal. Defenders try to get the ball away from the other team. Forwards try to kick the ball into the other team's goal. Games usually last ninety minutes. The team with the most points is the winner!

The team with the most points wins.

Ok. I think I understand now. In a team sport like soccer, each team wants to win. Players want to kick the ball so it is in the other team's goal in order to score points. They want to score points in order for their team to win.

Is that right?

Yes No

The system tries to generalize what it has learned.

I found some other team sports: basketball, football, and hockey. In these other team sports, do “players want to kick so that the ball is inside the other team's goal?”

Yes No

Then, it returns to the harder story, reads it and maps out an understanding, from temporal,to spatial reasoning.

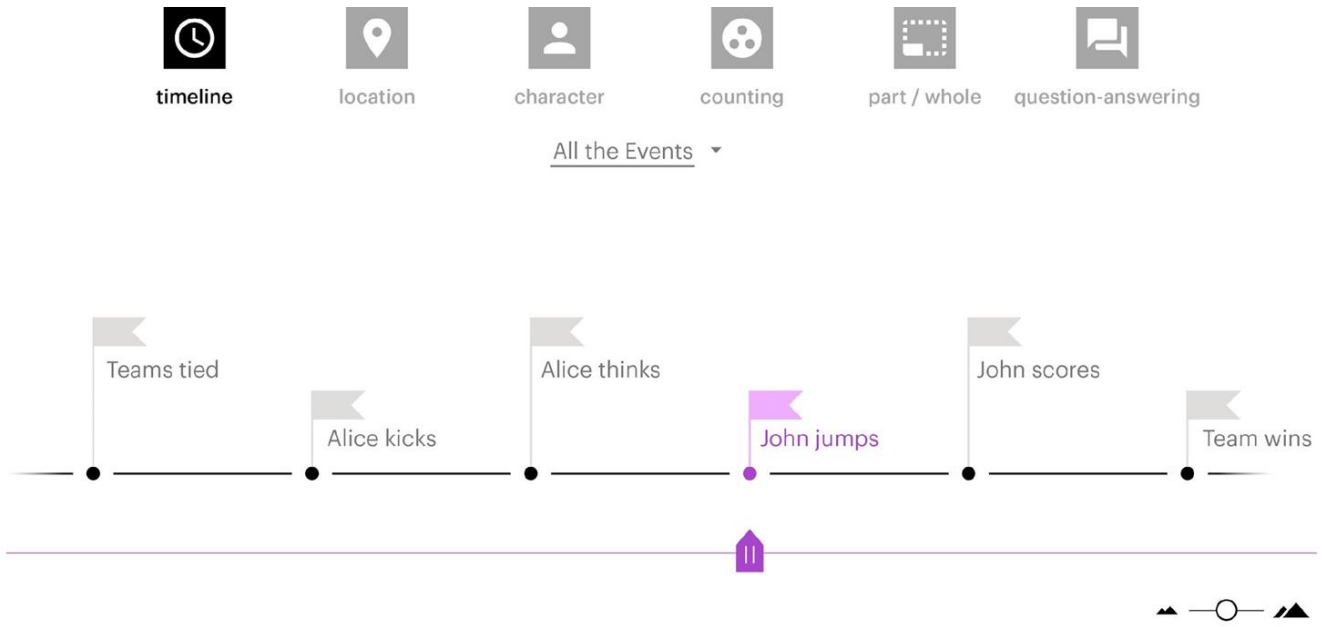
The Winning Shot

The soccer game was nearly over. The two teams were tied, one to one.

Alice kicked the ball. Oops! She kicked it the wrong way.

"What a bad kick," Alice thought.

But wait! Her teammate John jumped to stop the ball. The ball went off John and into the goal! Alice's team won the game. Hooray! It was not a bad kick, after all.



With that understanding, our system can answer and provide explanations.

The Winning Shot

The soccer game was nearly over. The two teams were tied, one to one.

Alice kicked the ball. Oops! She kicked it the wrong way.

"What a bad kick," Alice thought.

But wait! Her teammate John jumped to stop the ball. The ball went off John and into the goal!

X ~~Alice's team won the game.~~ Hooray! It was not a bad kick, after all.

✓ Correct

Explanation: It says, "Alice's team won the game." I don't know, otherwise.

✕ Incorrect

1 2 3 4 5

Test 2 | Question 1 of 5

Current Test Score: **40%**

Who won the game?

Answer: Alice's team.

✓ **Correct**

Explanation: Alice's team won because Alice's team had more points than the other team.

✓ **Correct**

1 2 3 4 5

With that understanding, our system can answer and provide explanations to any type of question.

The Winning Shot

The soccer game was nearly over. The two teams were tied, one to one.

Alice kicked the ball. Oops! She kicked it the wrong way.

"What a bad kick," Alice thought.

But wait! Her teammate John jumped to stop the ball. The ball went off John and into the goal! Alice's team won the game. Hooray! It was not a bad kick, after all.

Answer: Alice's team.
✓ Correct

Explanation: It says, "Alice's team won the game." I don't know, otherwise.
✗ Incorrect

1 2 3 4 5

Test 2 | Question 2 of 5

Current Test Score: **40%**

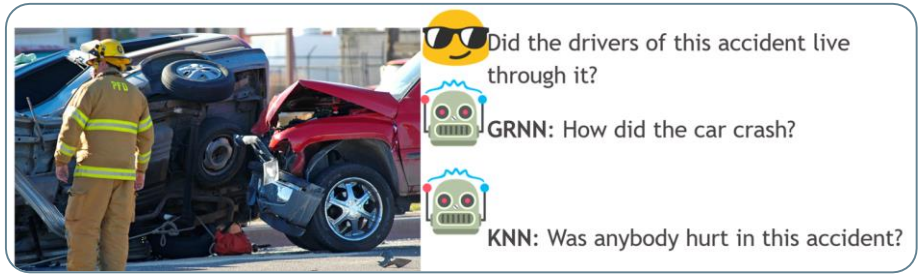
How many points did Alice's team have?

Answer: 2.
✓ Correct

Explanation: First, Alice's team had 1 point. Then John moved the ball into the goal. John is part of Alice's team. Therefore, Alice's team had 2 points.
✓ Correct

1 2 3 4 5

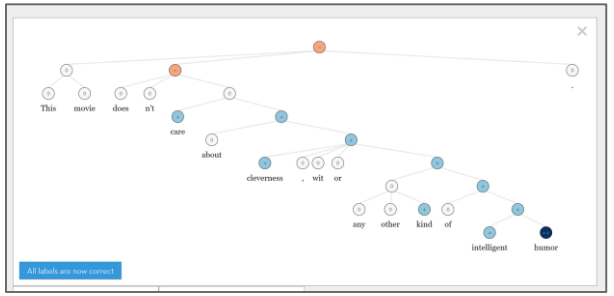
Discussion



Visual Question Generation

Temporal
Reading Comprehension
Event
Time
Causality
Reasoning
Knowledge Acquisition
Language Understanding
Common-sense
Natural Language Inference
Question Answering

Narrative Structure Learning
Story Generation
Story Understanding
Knowledge Base Completion



Sentiment Analysis

Visual Storytelling

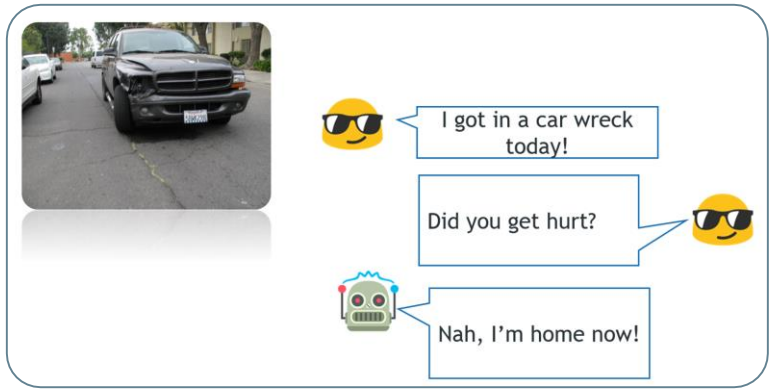
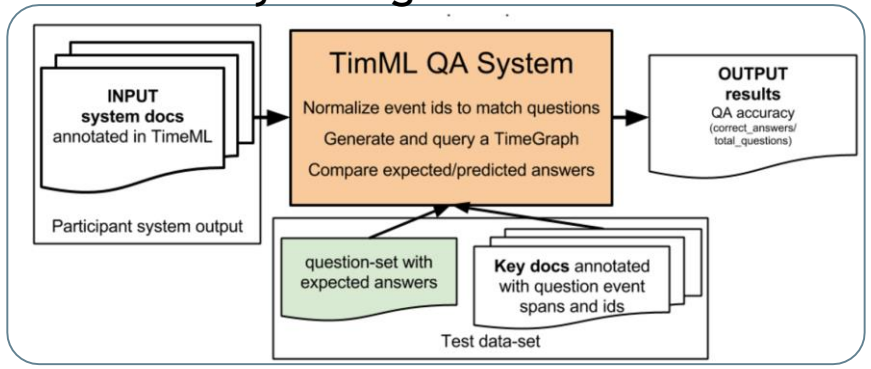


Image-Grounded Conversations



Temporal Question Answering

Context: John spends $20 a day on pickles. He decides to make his own to save money. He puts the pickles in brine. John waits 2 weeks for his pickles to get sour.

1: Now he is so happy that he has money.

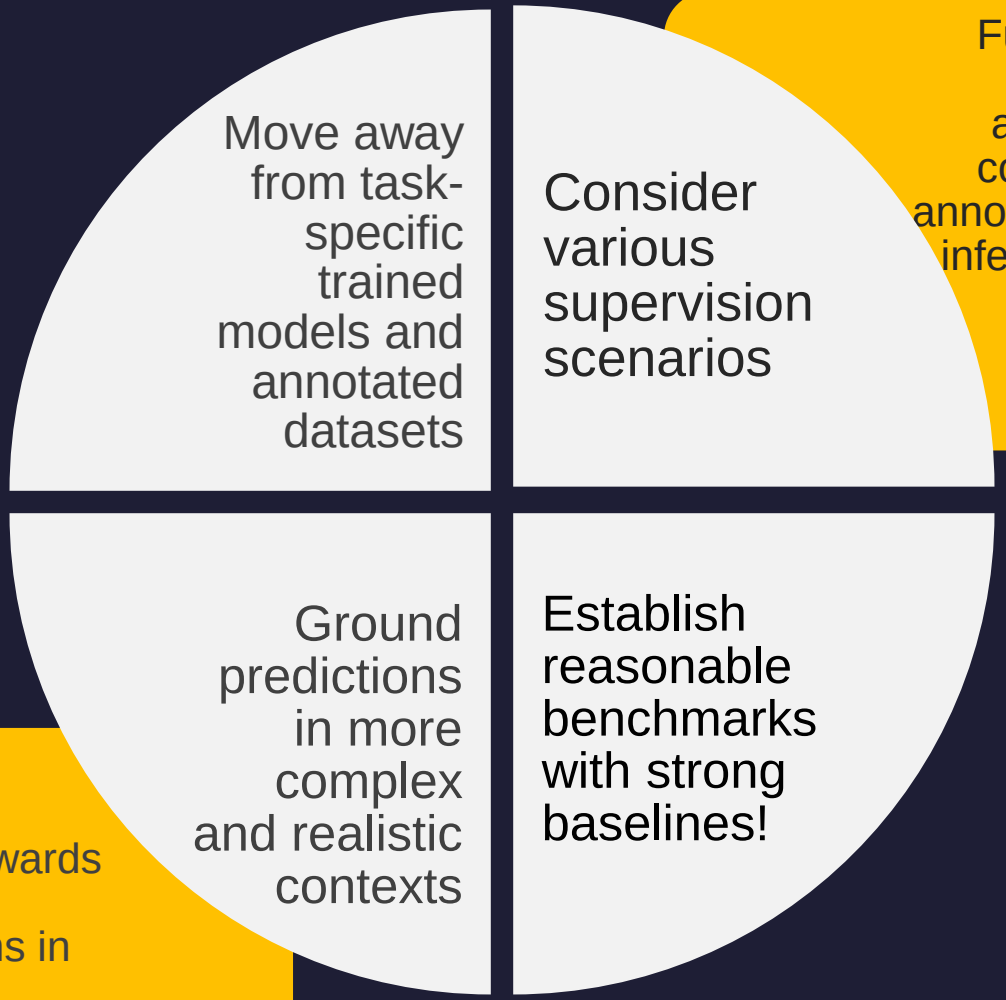
Story Understanding & Story Generation

The Current Trend in AI and NLP

For a particular narrow task:

Repeat for a new task!

What can we do?



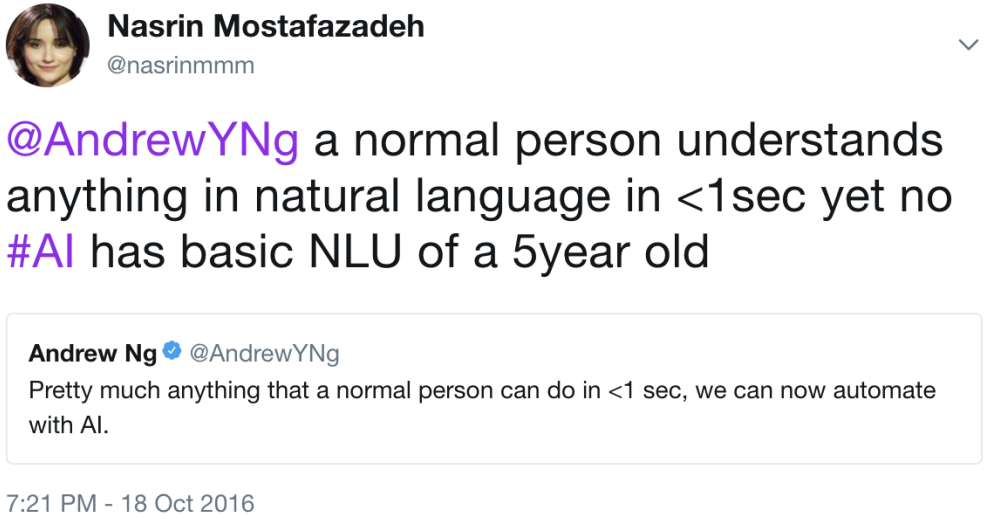
- Fully supervised models are not applicable when collecting a large annotated dataset is infeasible, which is often the case.

- + Move towards tangible applications in real world!

Conclusion: The State of NLU

- We've made a great progress in perception tasks such as 'speech recognition' and 'image recognition'

- Going beyond pattern recognition and tackling cognition, reasoning, and answering "why", is what remains extremely challenging and fundamentally unsolved today in AI.



Thanks to

James Allen, Rishi Sharma, Omid Bakhshandeh, David Ferruci, Lucy Vanderwende, Margaret Mitchell, Nathanael Chambers, Pushmeet Kohli, Chris Brockett, Bill Dolan, Michel Galley, Xiaodong He, Devi Parikh, Dhruv Batra, Ishan Misra, Jacob Devlin, Jianfeng Gao, Alyson Grealish, and many others …

Thanks for Listening ☺

Any Questions?