Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 7.2: Multimodal Inference and Knowledge

Paul Liang

** Co-lecturer: Louis-Philippe Morency. Original course co-developed with Tadas Baltrusaitis. Spring 2021 and 2022 editions taught by Yonatan Bisk. Spring 2023 edition taught by Yonatan and Daniel Fried*

Midterm Project Report Instructions

- **Goal:** Evaluate state-of-the-art models on your dataset and identify key issues through a detailed error analysis
  - It will inform the design of your new research ideas
- **Report format:** 2 column (ICML template)
  - The report should follow a similar structure to a research paper
  - Teams of 3: 8 pages, Teams of 4: 9 pages, Teams of 5: 10 pages.
- **Number of SOTA baseline models**
  - Teams of N should have at least N-1 baseline models
- **Error analysis**
  - This is one of the most important part of this report. You need to understand where previous models can be improved.

Examples of Possible Error Analysis Approaches

- Dataset-based:
  - Split correct/incorrect by label
  - Manually inspect the samples that are incorrectly predicted
    - What are the commonalities?
    - What are differences with the correct ones?
  - Sub-dataset analysis: length of question, rare words, cluttered images, high frequency in signals?

Examples of Possible Error Analysis Approaches

- Perturbation-based:
  - Make targeted changes to specific parts of the image.
  - Change one word/paraphrase/add redundant tokens.
  - See whether the model remains robust

Examples of Possible Error Analysis Approaches

- Model-based:
  - Visualize feature attributions: LIME, $1^{st}/2^{nd}$ order gradients
  - Ablation studies to understand what model components are important
- Theory-based:
  - Write out the math! From optimization and learning perspective, does the model do what's expected?
  - Some useful tools: consider linear case/other simplest case and derive solution, do empirical sanity checks first.

Examples of Possible Error Analysis Approaches

Published as a conference paper at ICLR 2018

On the Convergence of Adam and Beyond

Sashank J. Reddi, Satyen Kale & Sanjiv Kumar
Google New York
New York, NY 10011, USA
{sashank,satyenkale,sanjivk}@google.com

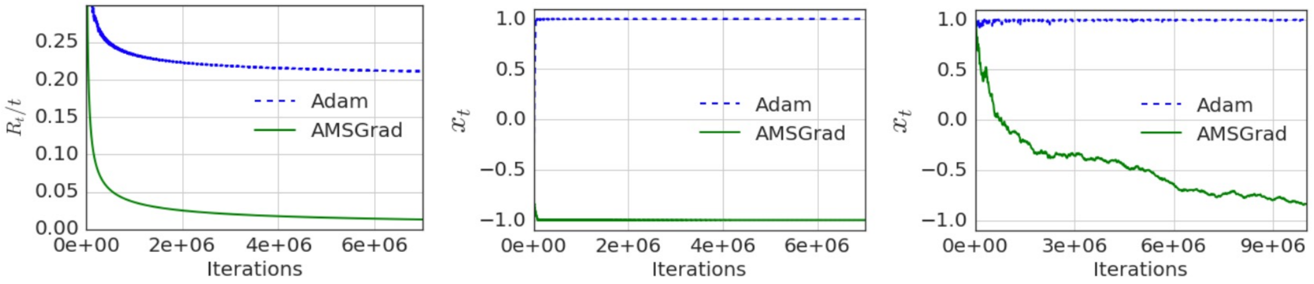
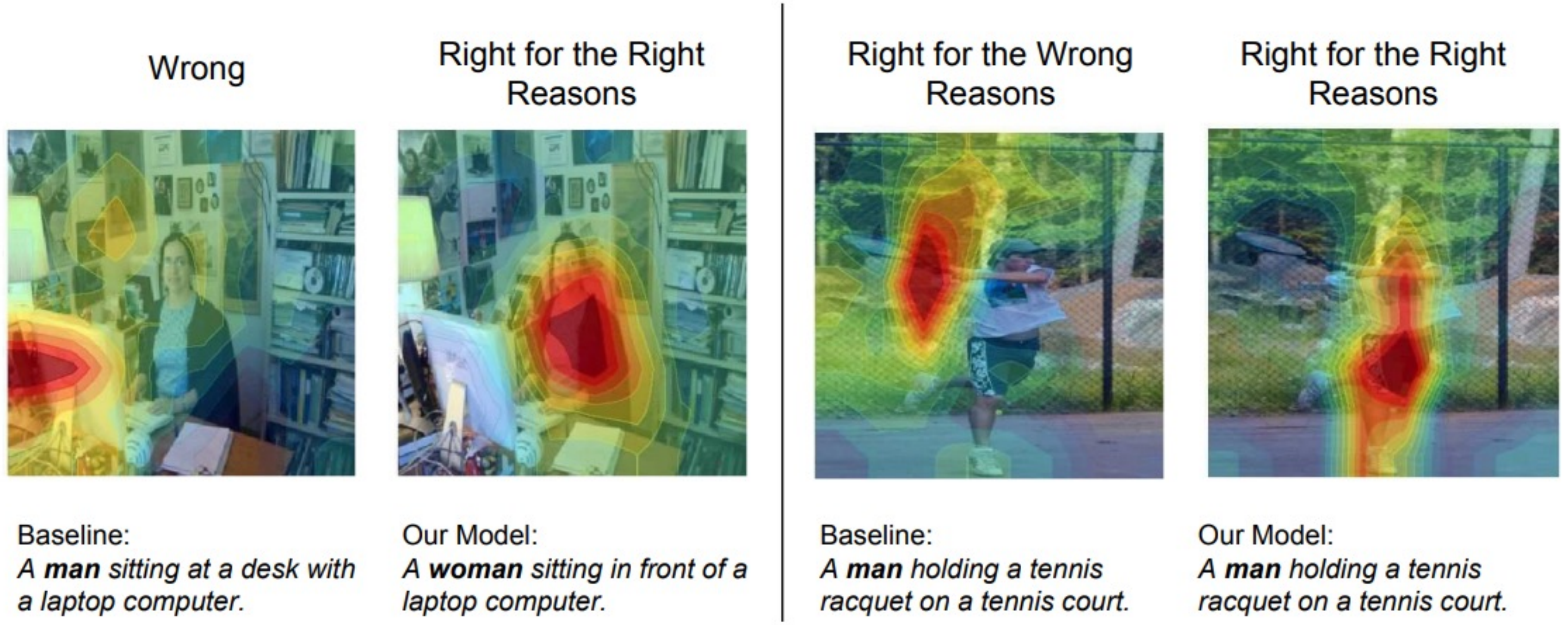


Figure 1: Performance comparison of Adam and AMSGrad on synthetic example on a simple one dimensional convex problem inspired by our examples of non-convergence. The first two plots (left and center) are for the online setting and the the last one (right) is for the stochastic setting.

[Reddi et al., On the Convergence of Adam and Beyond. ICLR 2018]

Examples of Possible Error Analysis Approaches

Finding: Image captioning models capture spurious correlations between gender and generated actions



You'll see more in today's reasoning lecture and in quantification lectures

[Hendricks et al., Women also Snowboard: Overcoming Bias in Captioning Models. ECCV 2018]

Midterm Project Report Instructions

Main report sections:

- Abstract
- Introduction
- Related work
- Problem statement
- Multimodal baseline models
- Experimental methodology
- Results and discussion
- New research ideas

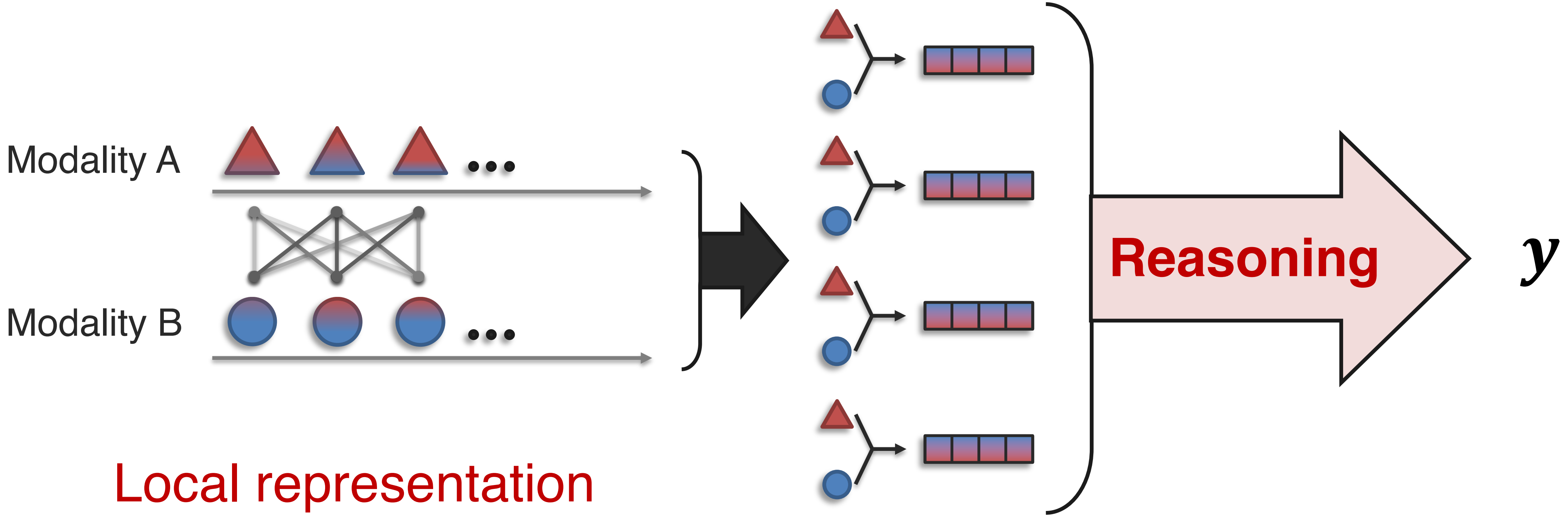
The structure is similar to a research paper submission ☺

Upcoming Deadlines

- Sunday October 29 8pm: Midterm report deadline
- Tuesday and Thursday (10/31 and 11/2): midterm presentations
  - All students are expected to attend both presentation sessions in person
  - Each team will present either Tuesday or Thursday
  - The focus of these presentations is about your research ideas
  - Feedback will be given by all students, instructors and TAs

Reasoning

Definition: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure.



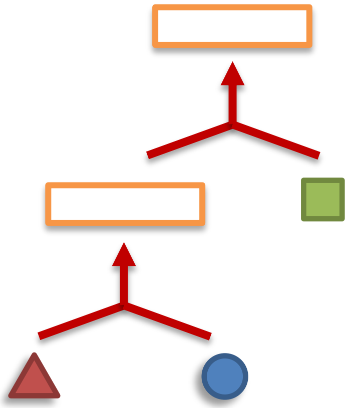
Local representation

+ Aligned representation

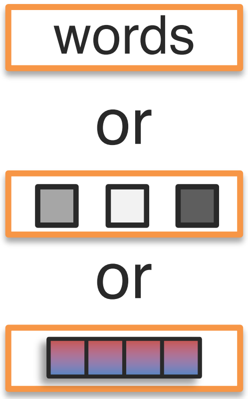
Reasoning

Definition: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure.

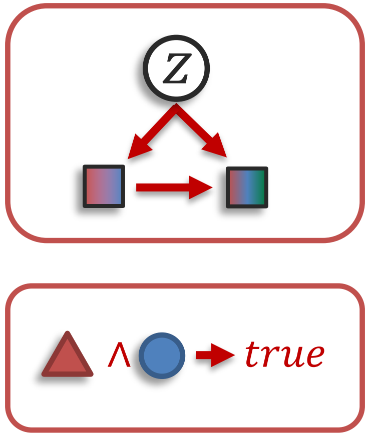
(A) Structure modeling



(B) Intermediate concepts



(C) Inference paradigm



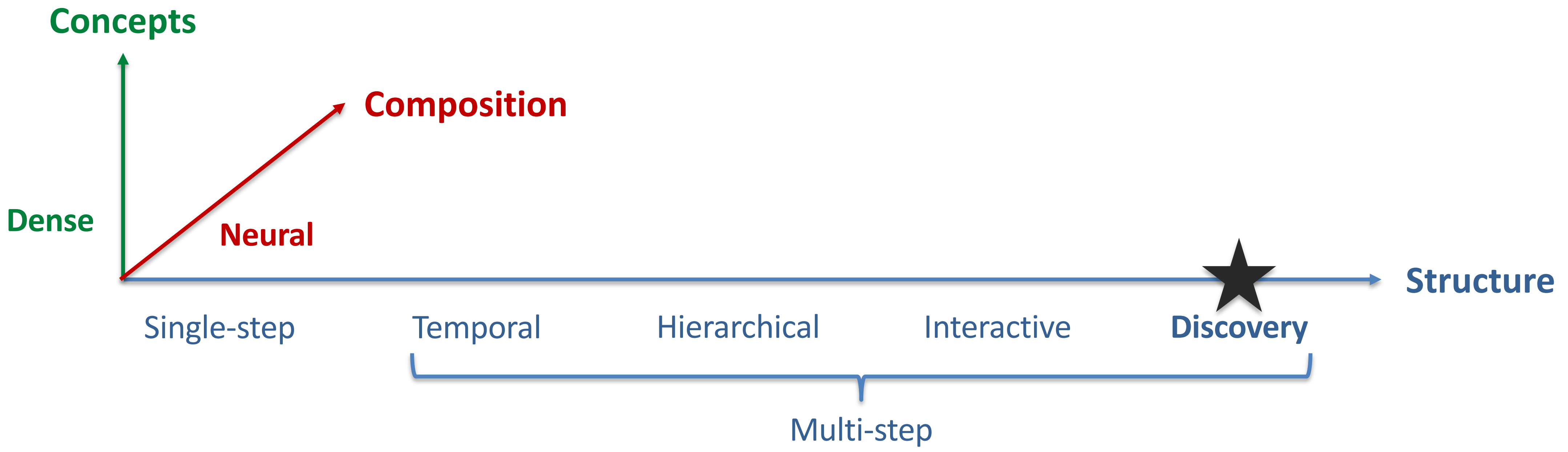
(D) External knowledge

Summary

Definition: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure.

	A Structure modeling	B Intermediate concepts	C Inference paradigm	D External knowledge
Last Thursday	Temporal Hierarchical	Continuous		
Tuesday	Interactive			
Today	Discovery	Discrete	Causal Logical	Knowledge Commonsense

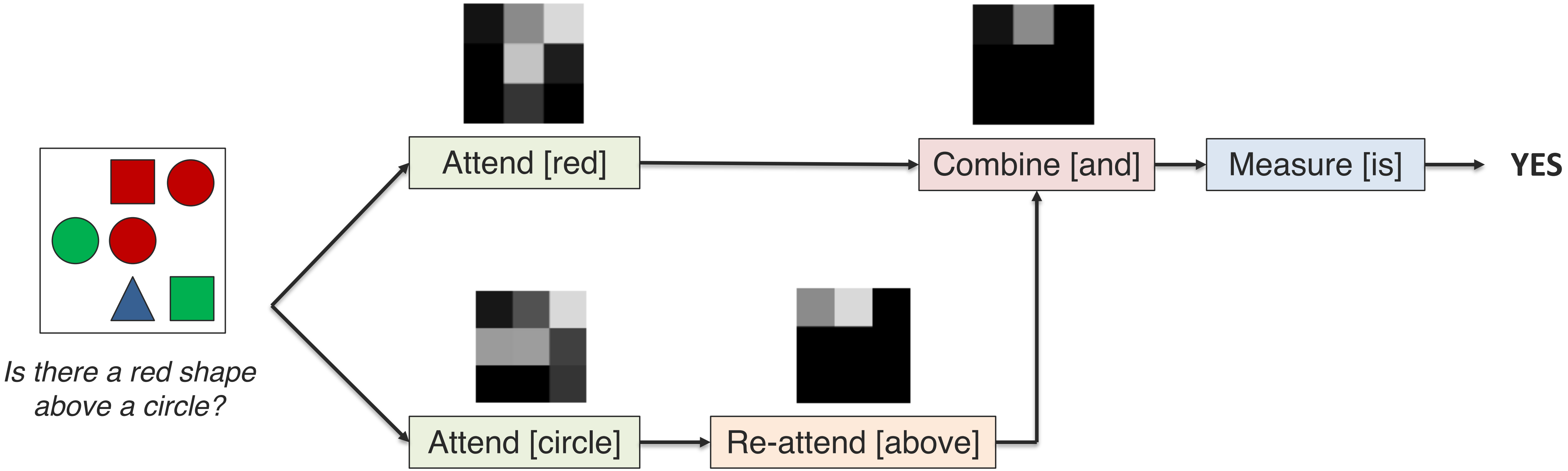
Sub-Challenge 3a: Structure Modeling



Structure Discovery

End-to-end neural module networks

Recall structure - leverage syntactic structure of language based on parsing



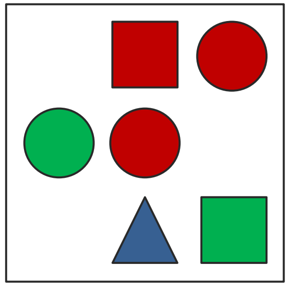
[Andreas et al., Neural Module Networks. CVPR 2016]

Structure Discovery

End-to-end neural module networks

Can we learn the structure end-to-end?

$$\tau \sim \pi_\theta(a|s)$$



Is there a red shape above a circle?

$\pi_\theta(a|s)$ → **?** →

- Attend [red]
- Attend [circle]
- Re-attend [above]
- Combine [and]
- Measure [is]

→ **NMN** → **YES**

$r(\tau)$

[Hu et al., Learning to Reason: End-to-End Module Networks for Visual Question Answering. ICCV 2017]

Stochastic Optimization

$$\max_{\theta} \mathbb{E}_{q_\theta(\mathbf{z})}[f(\mathbf{z})]$$

RL

$$\max_{\theta} J(\theta)$$ Reward

$$\max_{\theta} \mathbb{E}_{\tau \sim p(\tau;\theta)}[r(\tau)]$$

In RL (at least for discrete actions):
- T is a sequence of discrete actions
- p(T; θ) is not reparameterizable
- r(T) is a black box function
i.e. the environment

$$s \xrightarrow{\pi_\theta(a|s)} a \xrightarrow{???} r$$

REINFORCE is a general-purpose solution!

Revisiting REINFORCE

$$\max_{\theta} \mathbb{E}_{q_\theta(\mathbf{z})}[f(\mathbf{z})]$$

(we will revisit this equation for generative models)

We want to take gradients wrt θ of the term:

$$\nabla_\theta \mathbb{E}_{q_\theta(\mathbf{z})}[f(\mathbf{z})] = \mathbb{E}_{q_\theta(\mathbf{z})}[f(\mathbf{z}) \nabla_\theta \log q_\theta(\mathbf{z})]$$

We can now compute a Monte Carlo estimate:

Sample $\mathbf{z}^1, \mathbf{z}^2, \ldots, \mathbf{z}^K$ from $q_\theta(\mathbf{z})$ and estimate

$$\nabla_\theta \mathbb{E}_{q_\theta(\mathbf{z})}[f(\mathbf{z})] \approx \frac{1}{K} \sum_k [f(\mathbf{z}^k) \nabla_\theta \log q_\theta(\mathbf{z}^k)]$$

What we derived: sample trajectories and compute:

$$\boxed{\nabla_\theta J(\theta) \approx \sum_{t \geq 0} r(\tau) \nabla_\theta \log \pi_\theta(a_t | s_t)}$$

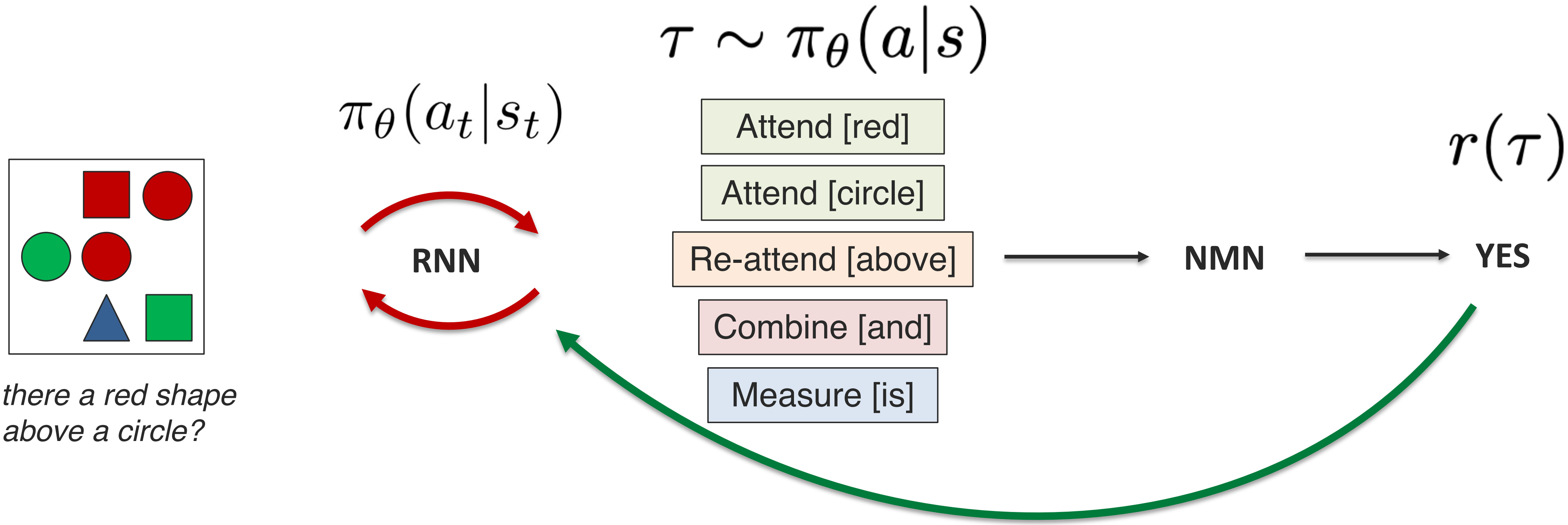
- z can be discrete or continuous!
- q(z) can be a discrete and continuous distribution!
- q(z) must allow for easy sampling and be differentiable wrt θ
- f(z) can be a black box!

Structure Discovery

End-to-end neural module networks

Can we learn the structure end-to-end?

$$\boxed{\nabla_\theta J(\theta) \approx \sum_{t \geq 0} r(\tau) \nabla_\theta \log \pi_\theta(a_t | s_t)}$$



Is there a red shape above a circle?

[Hu et al., Learning to Reason: End-to-End Module Networks for Visual Question Answering. ICCV 2017]

Structure Discovery

Structure fully learned from optimization and data

1. Define basic representation building blocks

ReLU | Layer norm | Conv | Self-attention

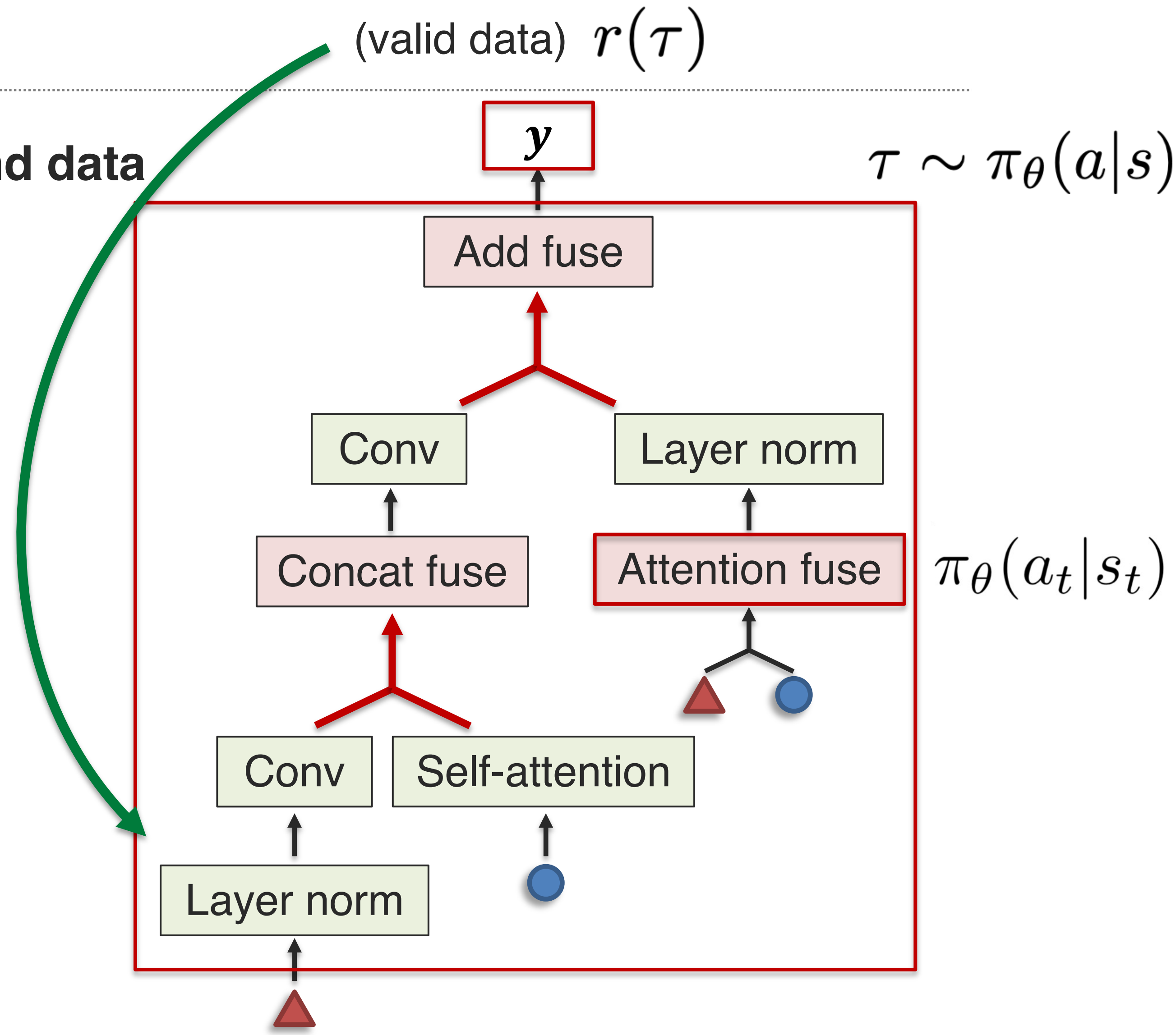
2. Define basic fusion building blocks

Concat fuse | Attention fuse | Add fuse

3. Automatically search for composition using neural architecture search

$$\nabla_\theta J(\theta) \approx \sum_{t \geq 0} r(\tau) \nabla_\theta \log \pi_\theta(a_t | s_t)$$

Nice, but slow!



[Xu et al., MUFASA: Multimodal Fusion Architecture Search for Electronic Health Records. AAAI 2021]

Continuous Structure Discovery

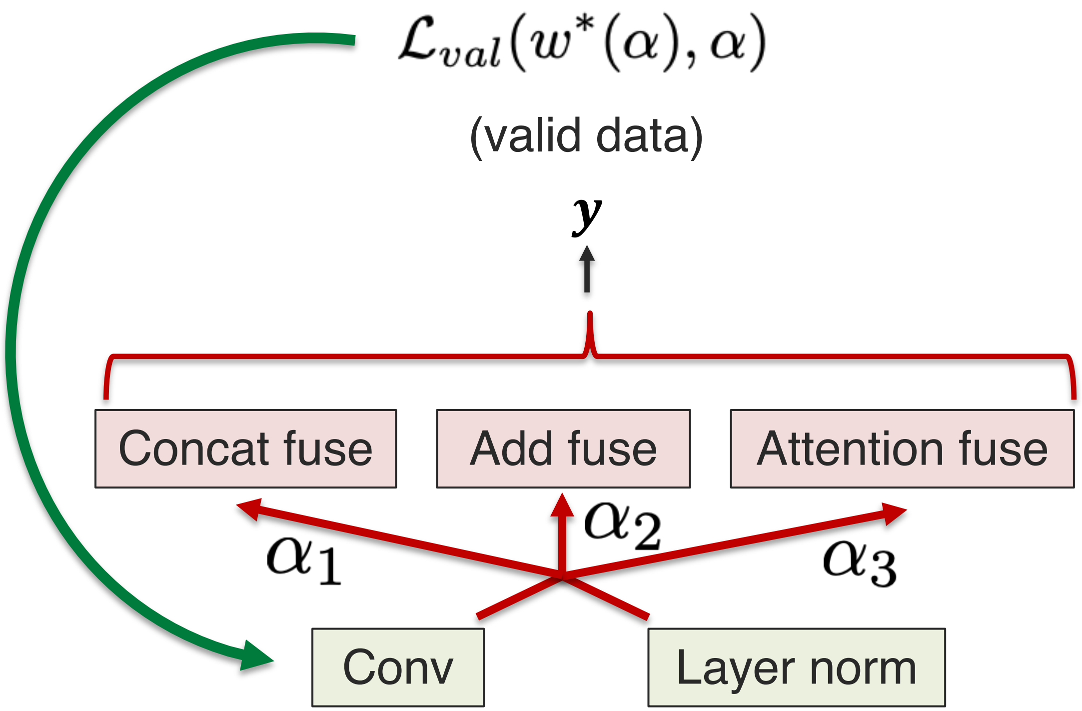
Biggest problem: discrete optimization is slow.
Differentiable optimization for structure learning:

1. Approximate selection with softmax:

$$o'(x) = \sum_i \frac{\exp(\alpha_i)}{\sum_i \exp(\alpha_i)} o_i(x)$$

2. Solve bi-level optimization problem

$$\min_{\alpha} \quad \mathcal{L}_{val}(w^*(\alpha), \alpha)$$
$$\text{s.t.} \quad w^*(\alpha) = \text{argmin}_w \ \mathcal{L}_{train}(w, \alpha)$$



[Liu et al., DARTS: Differentiable Architecture Search. ICLR 2019]

Continuous Structure Discovery

Biggest problem: discrete optimization is slow.
Differentiable optimization for structure learning:

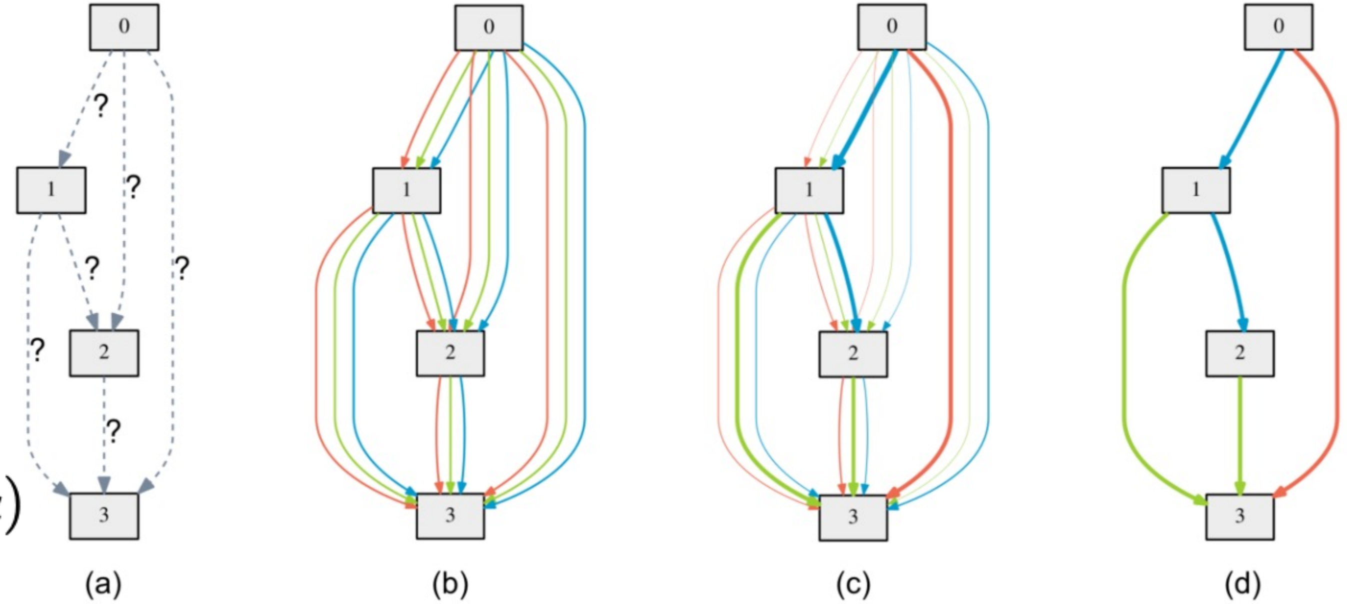
1. Approximate selection with softmax:

$$o'(x) = \sum_i \frac{\exp(\alpha_i)}{\sum_i \exp(\alpha_i)} o_i(x)$$

2. Solve bi-level optimization problem

$$\begin{aligned} \min_{\alpha} \quad & \mathcal{L}_{val}(w^*(\alpha), \alpha) \\ \text{s.t.} \quad & w^*(\alpha) = \text{argmin}_w \;\; \mathcal{L}_{train}(w, \alpha) \end{aligned}$$

3. Convert softmax to argmax



Faster but still non-trivial

[Liu et al., DARTS: Differentiable Architecture Search. ICLR 2019]

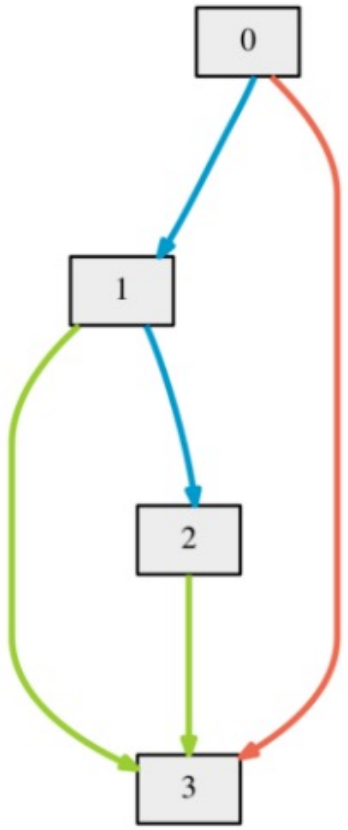
Continuous Structure Discovery

In general, optimization over directed acyclic graphs (DAGs):

Graph ***G***, Data ***X***, Adjacency matrix ***W:***

$$\min_{W} \ \ell(W; X) \quad \text{s.t.} \ G(W) \in DAG \qquad \overset{?}{\Longleftrightarrow} \qquad \min_{W} \ \ell(W; X) \quad \text{s.t.} \ h(W) = 0$$

(combinatorial 😱) (smooth 😎)



(d)

[Zheng et al., DAGs with NO TEARS: Continuous Optimization for Structure Learning. NeurIPS 2018]

Continuous Structure Discovery

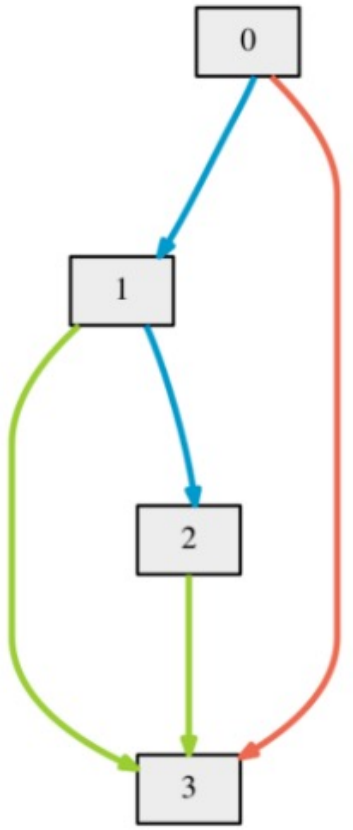
$$\min_W \ \ell(W;X) \quad \text{s.t.} \ G(W) \in DAG \quad \overset{?}{\Longleftrightarrow} \quad \min_W \ \ell(W;X) \quad \text{s.t.} \ h(W) = 0$$

$$h(W) = \operatorname{tr}(e^{W \circ W}) - d,$$

$$e^A = I + A + \frac{1}{2!}A^2 + \frac{1}{3!}A^3 + \cdots$$

- *K*-th power of adjacency matrix ***W*** counts the number of *k*-step paths from one node to another.
- If the diagonal of the matrix power is all zeros, there are no *k*-step cycles.
- Acyclic = check all *k = 1,2, …, size of graph.*

Can now do continuous optimization to solve for W, but **nonconvex**



(d)

[Zheng et al., DAGs with NO TEARS: Continuous Optimization for Structure Learning. NeurIPS 2018]

Continuous Structure Discovery

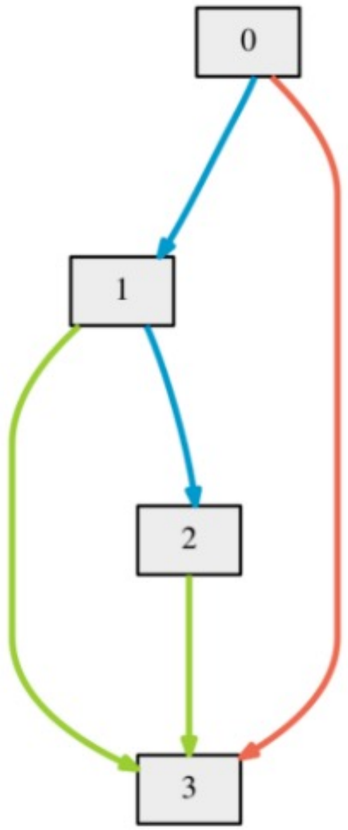
$$\min_{W} \ \ell(W;X) \quad \text{s.t.} \ G(W) \in DAG \quad \overset{?}{\Longleftrightarrow} \quad \min_{W} \ \ell(W;X) \quad \text{s.t.} \ h(W) = 0$$

In our paper, we showed that such a function h exists,

$$h(W) = \operatorname{tr}(e^{W \circ W}) - d,$$

and that it has a simple gradient:

$$\nabla h(W) = (e^{W \circ W})^T \circ 2W.$$

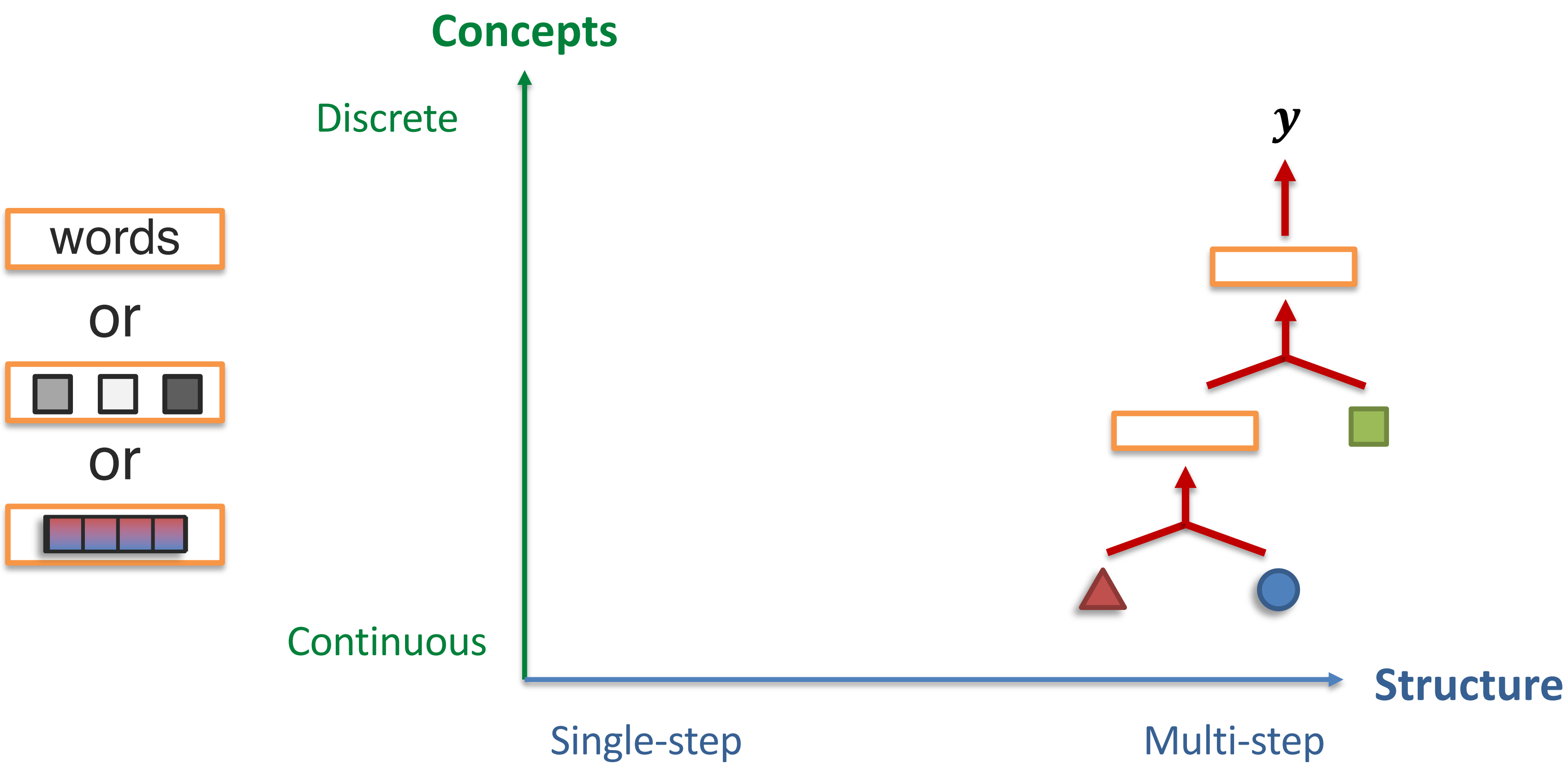


(d)

[Zheng et al., DAGs with NO TEARS: Continuous Optimization for Structure Learning. NeurIPS 2018]

Sub-Challenge 3b: Intermediate Concepts

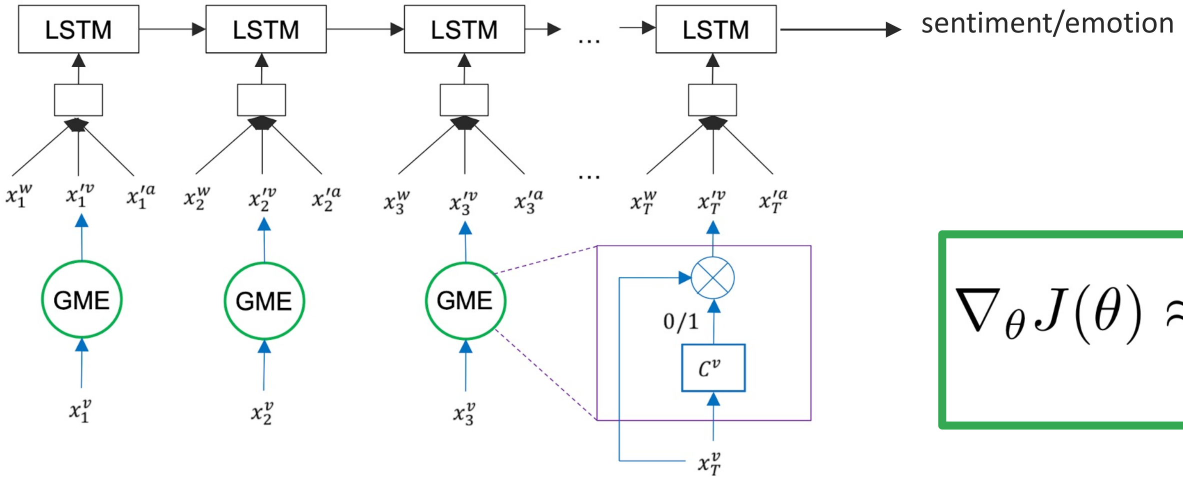
Definition: The parameterization of individual multimodal concepts in the reasoning process.



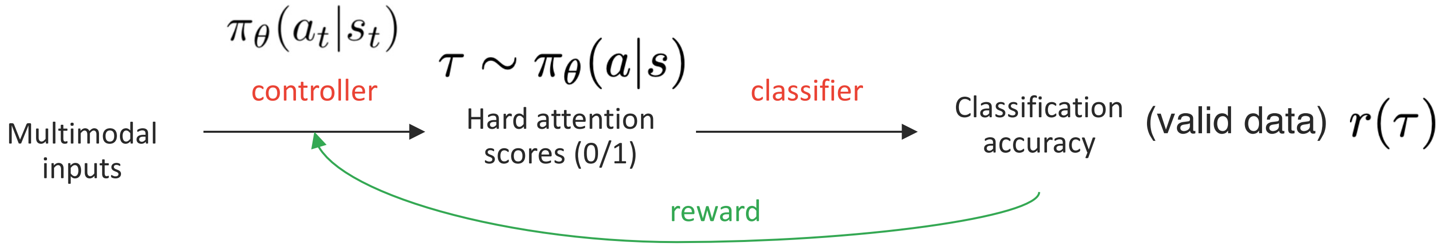
Discrete Concepts via Hard Attention

Hard attention 'gates' (0/1) rather than soft attention (softmax between 0-1)

- Can be seen as discrete layers in between differentiable neural net layers



$$\nabla_\theta J(\theta) \approx \sum_{t \geq 0} r(\tau) \nabla_\theta \log \pi_\theta(a_t | s_t)$$



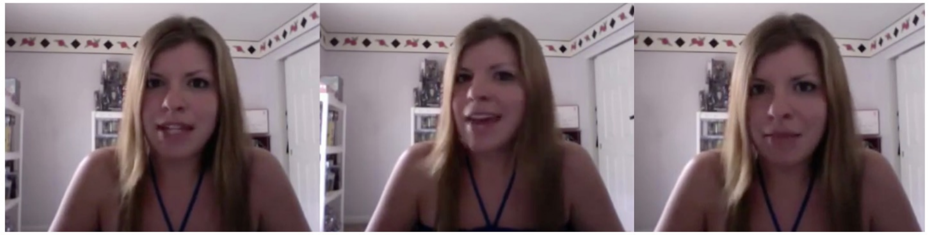
[Xu et al., Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. ICML 2015]
[Chen et al., Multimodal Sentiment Analysis with Word-level Fusion and Reinforcement Learning. ICMI 2017]

Discrete Concepts via Hard Attention

Hard attention 'gates' (0/1) rather than soft attention (softmax between 0-1)

- Can be seen as discrete layers in between differentiable neural net layers

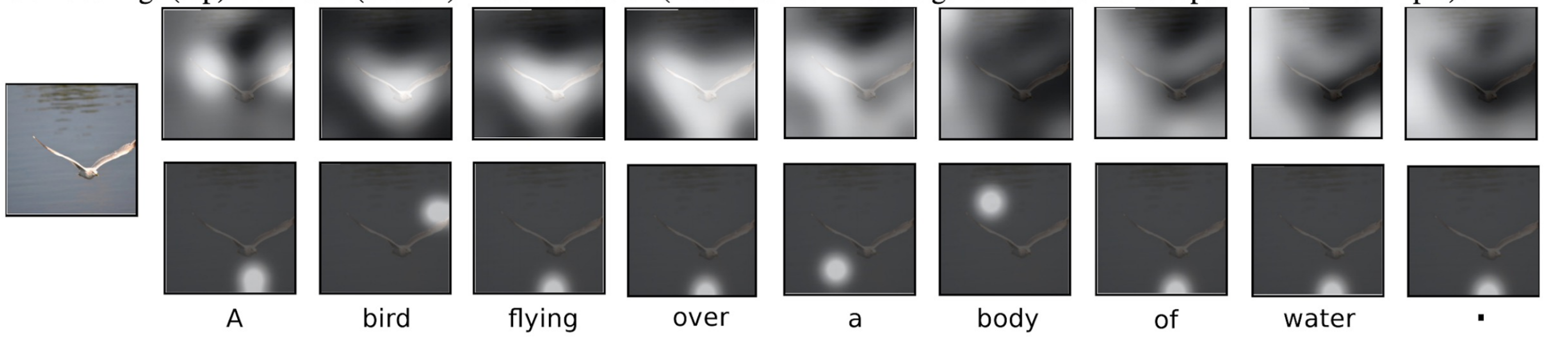
Sentiment analysis, emotion recognition



Reject Pass Reject

Image captioning

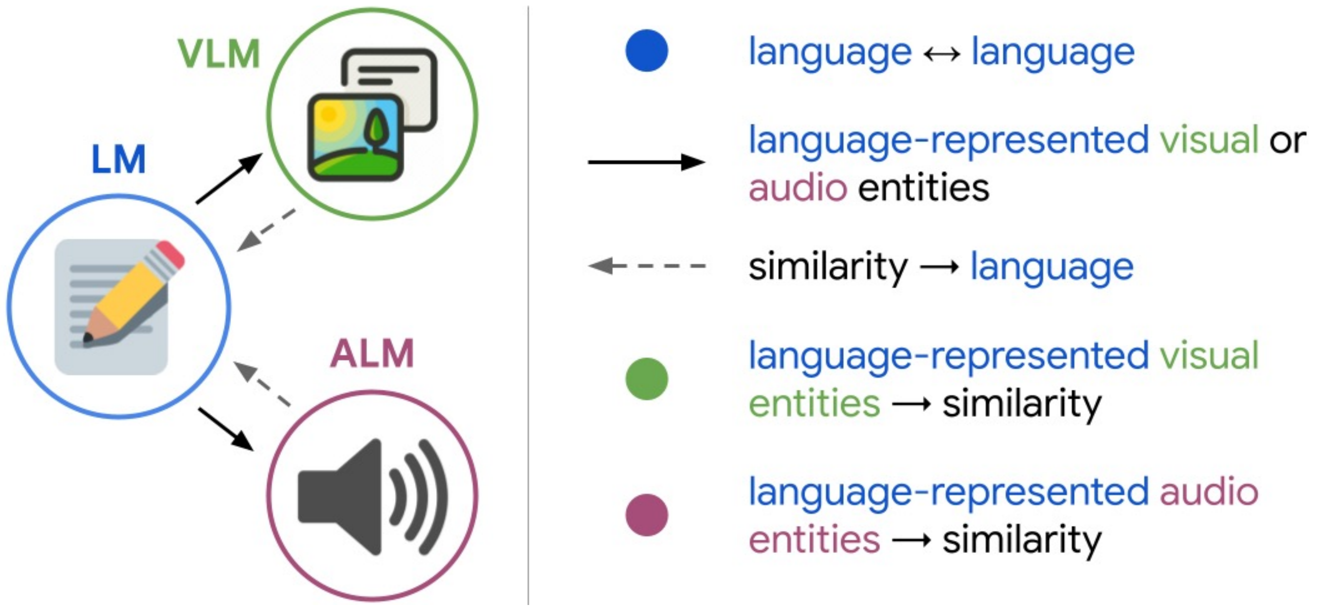
Figure 3. Visualization of the attention for each generated word. The rough visualizations obtained by upsampling the attention weights and smoothing. (top)"soft" and (bottom) "hard" attention (note that both models generated the same captions in this example).



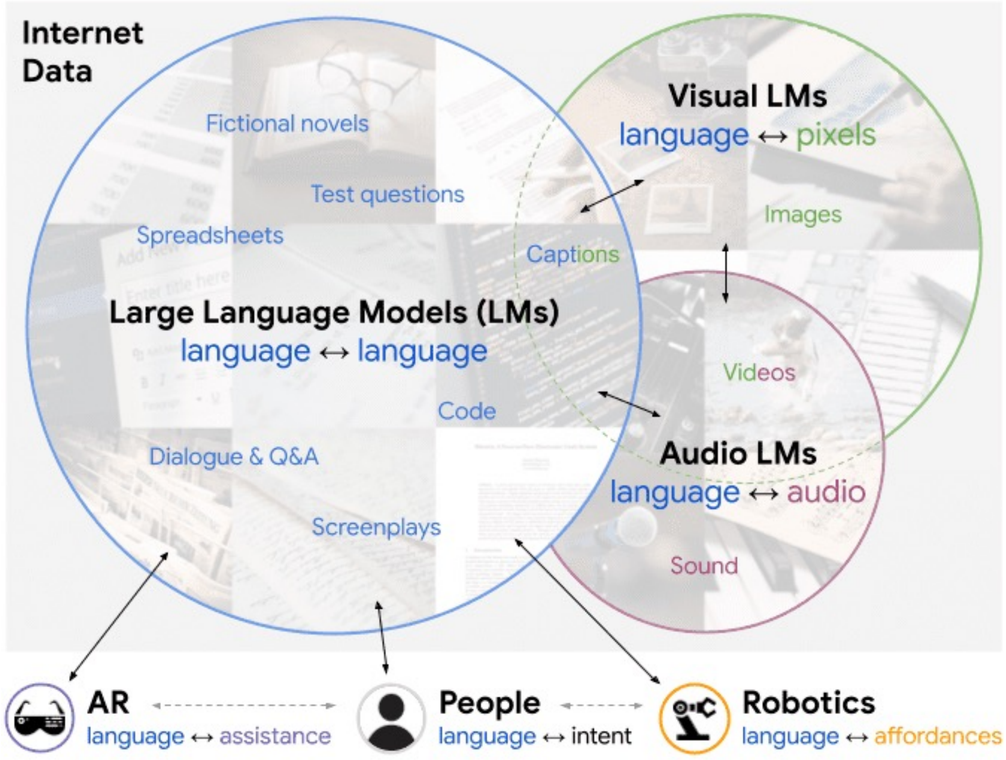
[Xu et al., Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. ICML 2015]
[Chen et al., Multimodal Sentiment Analysis with Word-level Fusion and Reinforcement Learning. ICMI 2017]

Discrete Concepts via Language

- Large language/video/audio models interacting with each other
- Each language model has its own distinct *domain knowledge*
- Interaction is scripted and zero-shot



Guided multimodal discussion



Combining domain knowledge

[Zeng et al., Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language. arXiv 2022]

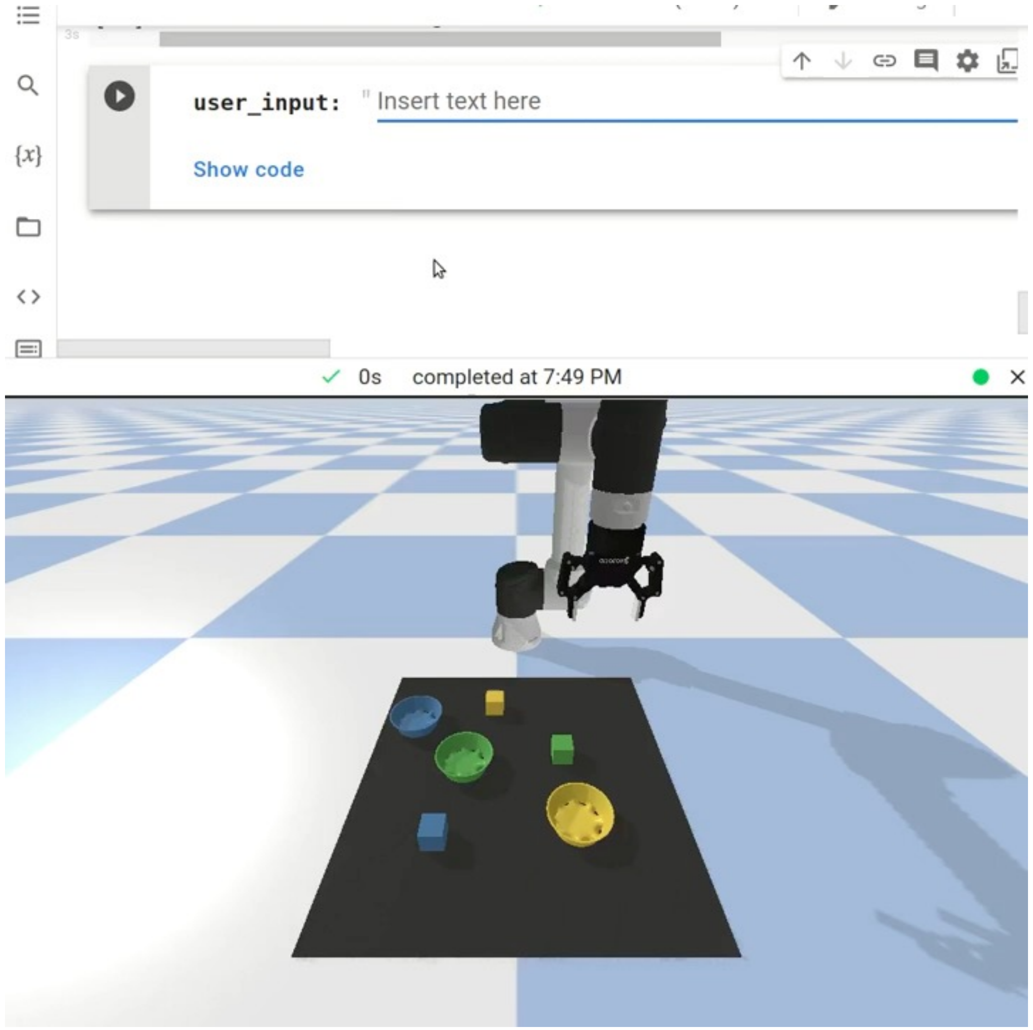
Discrete Concepts via Language

Image captioning

[Zeng et al., Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language. arXiv 2022]

Discrete Concepts via Language

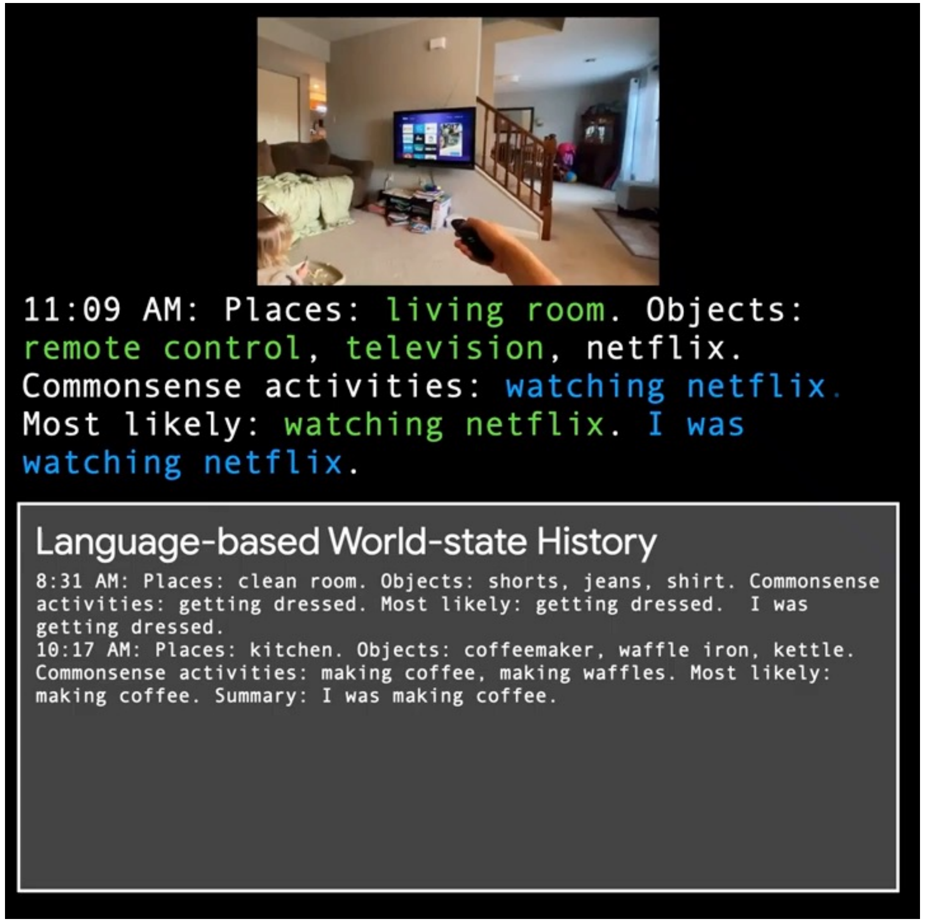
Robot perception and planning

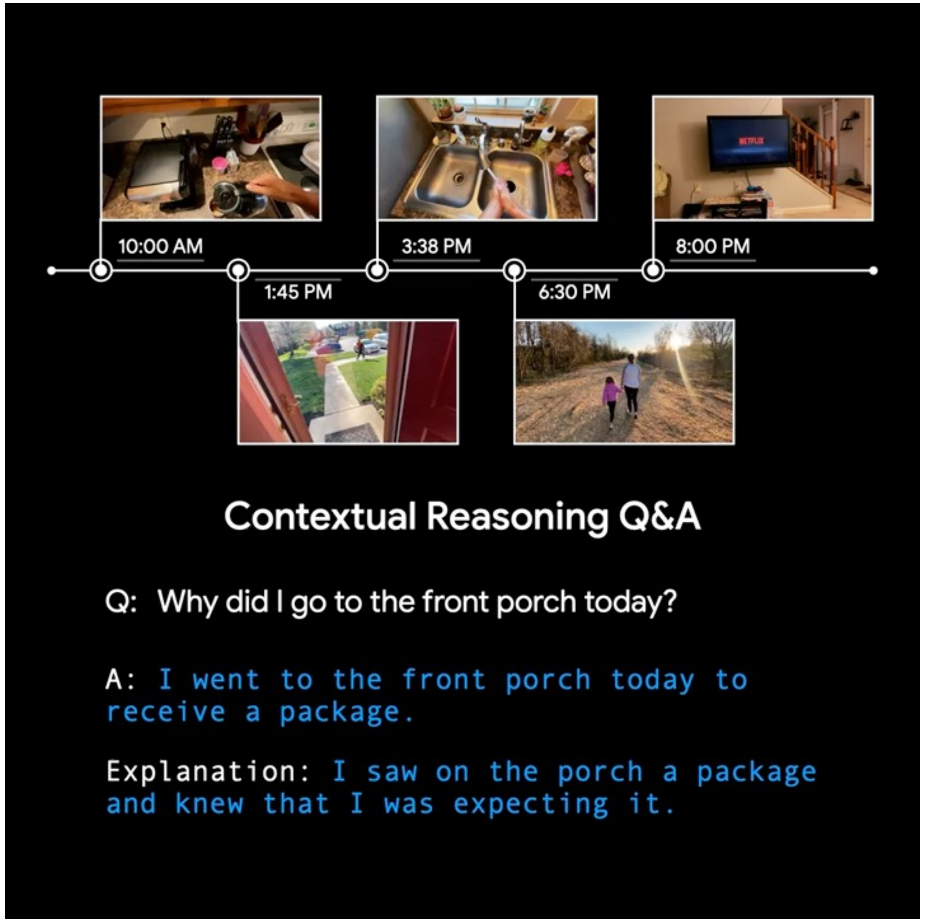


[Zeng et al., Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language. arXiv 2022]

Discrete Concepts via Language

Video reasoning



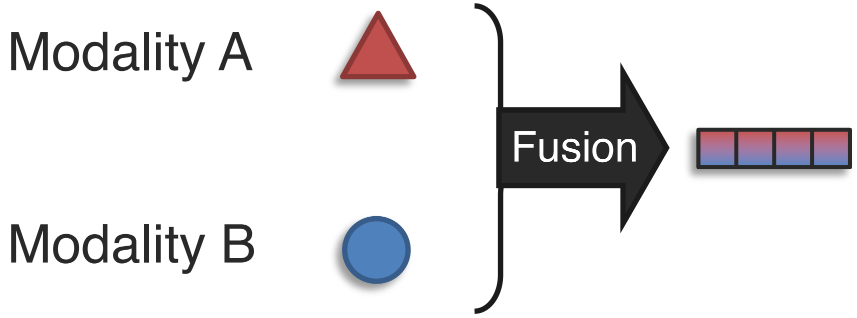


[Zeng et al., Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language. arXiv 2022]

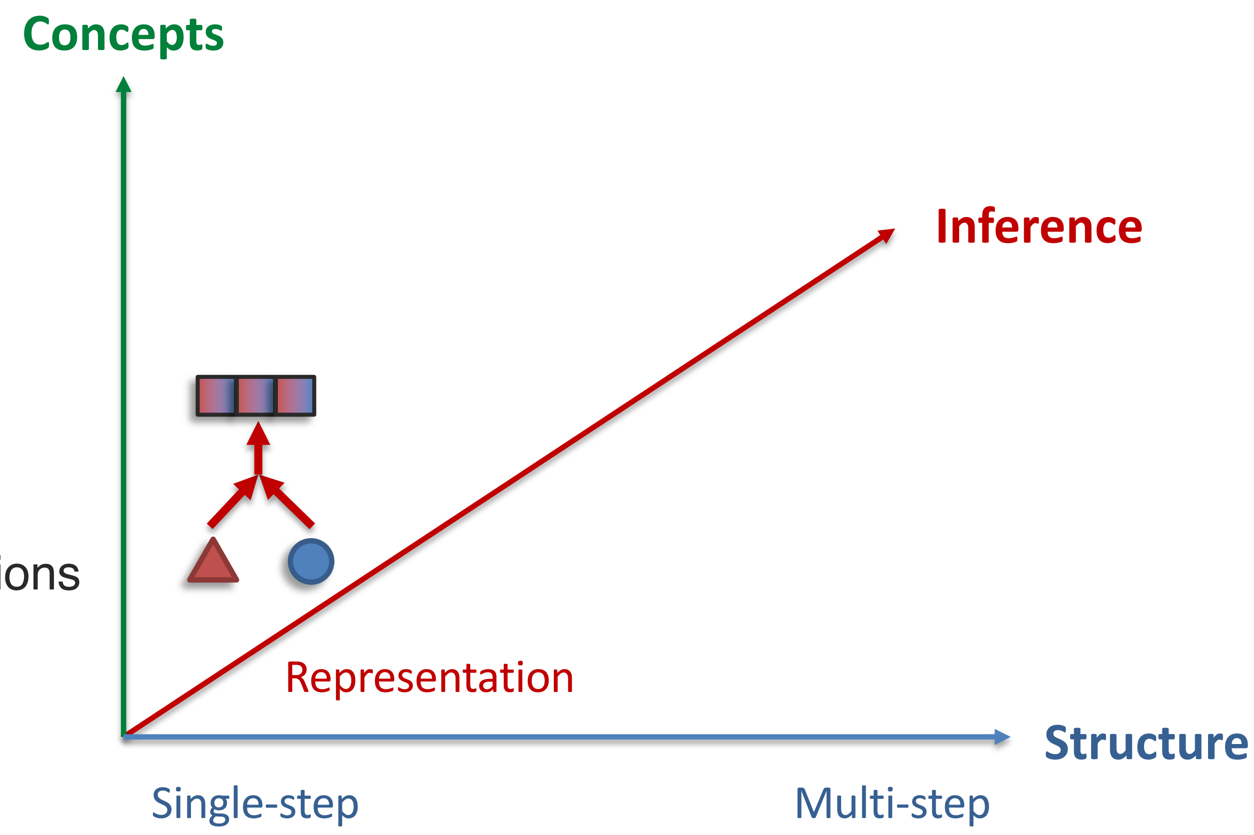
Sub-Challenge 3c: Inference Paradigm

Definition: How increasingly abstract concepts are inferred from individual multimodal evidences.

Recall representation fusion:



Potential issues:
- Models may capture spurious correlations
- Not robust to targeted manipulations
- Lack of interpretability/control

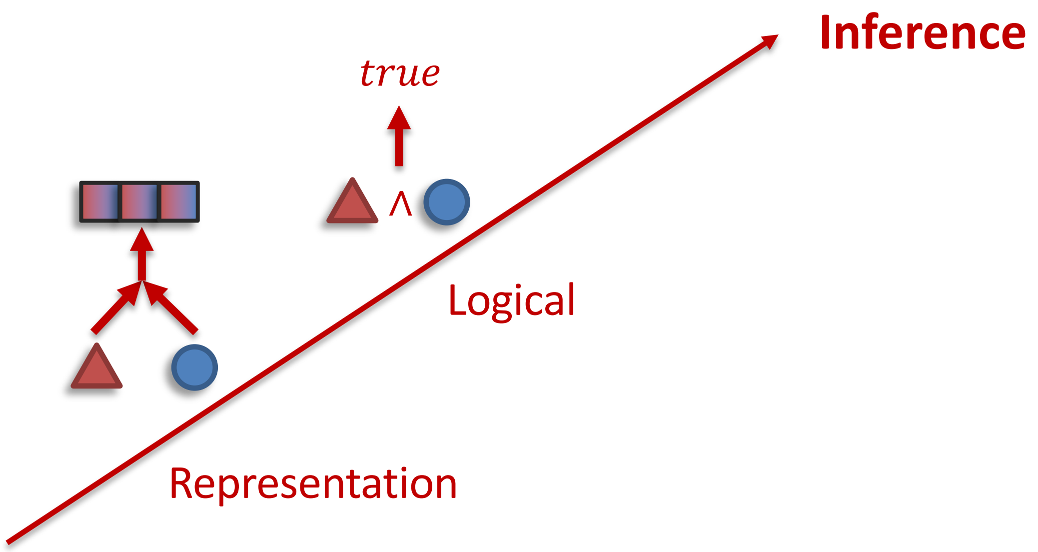


Sub-Challenge 3c: Inference Paradigm

Definition: How increasingly abstract concepts are inferred from individual multimodal evidences.

Towards explicit inference paradigms:

1. Logical inference: given premises inferred from multimodal evidence, how can one derive **logical** conclusions?



Logical Inference

Recall error analysis!

Inference through logical operators in question

Is there beer AND is there a WINE GLASS?

Is the man NOT wearing shoes AND is there beer?

Is there beer? *Is the man wearing shoes?*

Adversarial antonyms ✗

Logical connectives ✗

Basic premises ✓

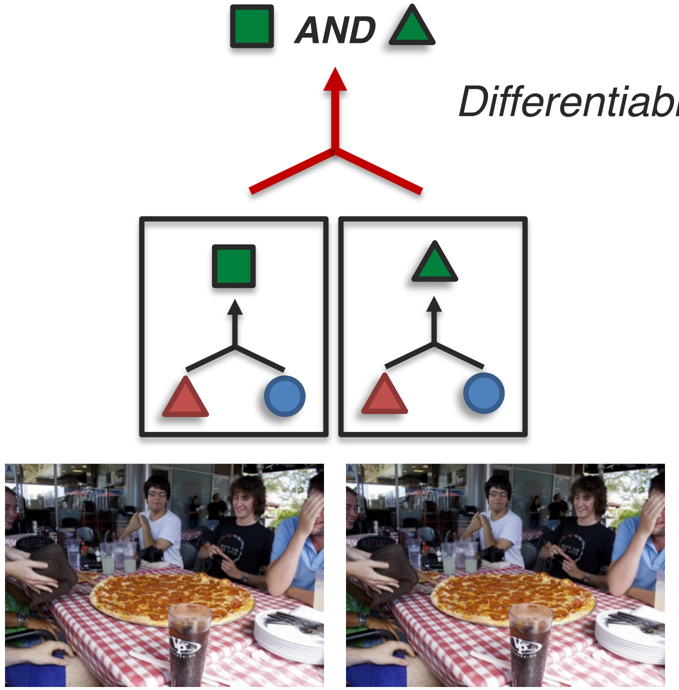
Existing models struggle to capture logical connectives.
How can we make them more logical?

[Gokhale et al., VQA-LOL: Visual Question Answering Under the Lens of Logic. ECCV 2020]

Logical Inference

Inference through logical operators in question

*Are they in a restaurant **AND** are they all boys?*



*Differentiable **AND** composition operator!*

Are they in a restaurant?

Are they all boys?

Also applies to other logic connectives: *AND, OR, NOT*

[Gokhale et al., VQA-LOL: Visual Question Answering Under the Lens of Logic. ECCV 2020]

Soft Logical Operators

Inference through logical operators in question

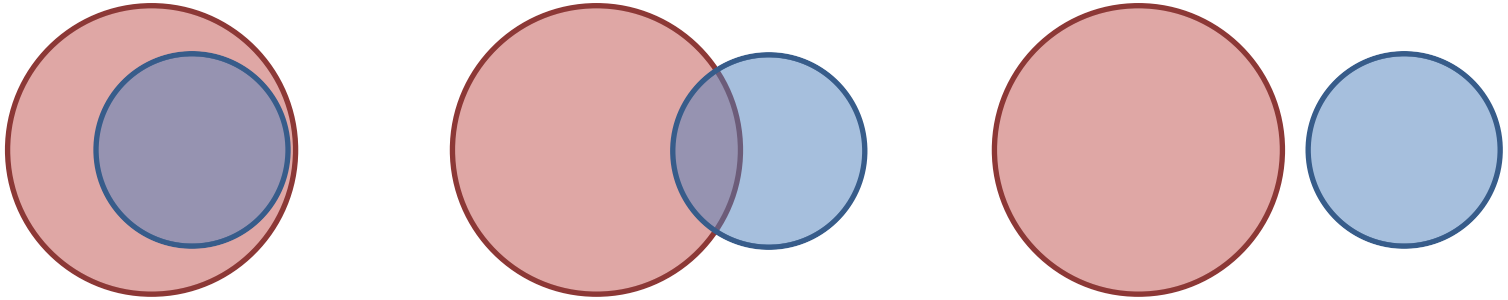
Fréchet inequalities to make logical functions differentiable:

- Probability of an intersection of events

$$\max(0,\, \mathbb{P}(A) + \mathbb{P}(B) - 1) \leq \mathbb{P}(A \cap B) \leq \min(\mathbb{P}(A),\, \mathbb{P}(B)),$$

- Probability of a union of events

$$\max(\mathbb{P}(A),\, \mathbb{P}(B)) \leq \mathbb{P}(A \cup B) \leq \min(1,\, \mathbb{P}(A) + \mathbb{P}(B)).$$

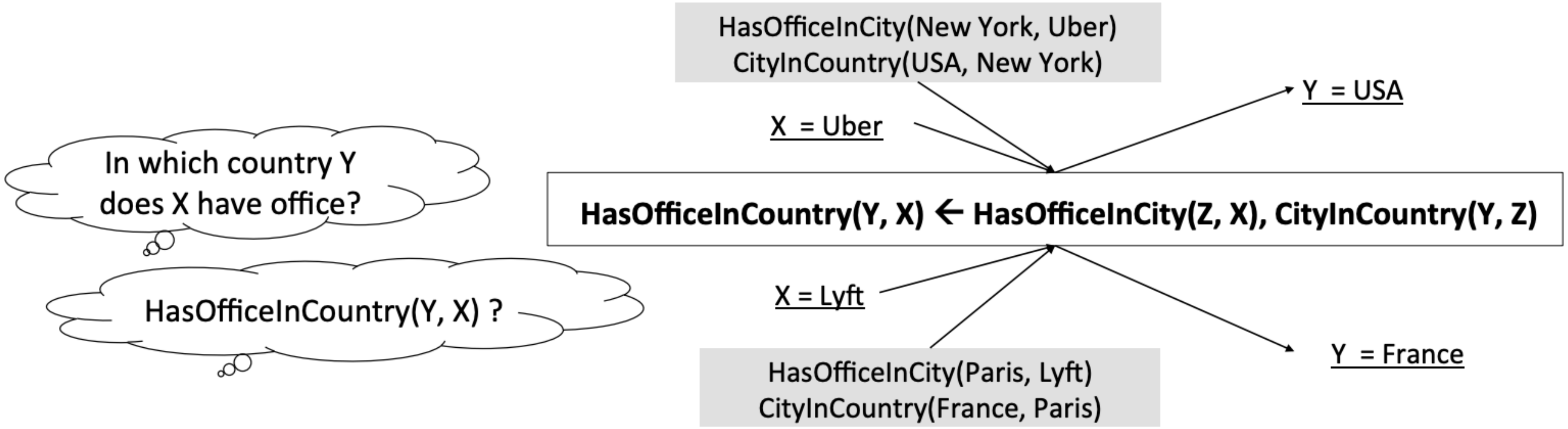


[Gokhale et al., VQA-LOL: Visual Question Answering Under the Lens of Logic. ECCV 2020]

Logical Inference Challenges

Open challenges

Many open directions



Differentiable knowledge base reasoning

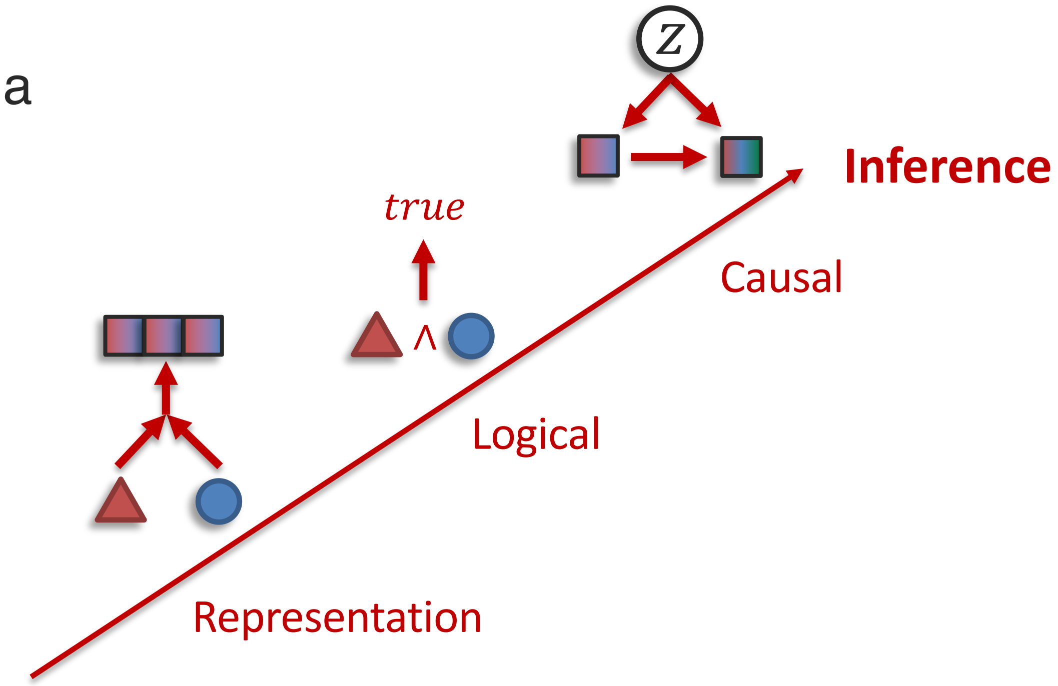
[Yang et al., Differentiable Learning of Logical Rules for Knowledge Base Reasoning. NeurIPS 2017]

Sub-Challenge 3c: Inference Paradigm

Definition: How increasingly abstract concepts are inferred from individual multimodal evidences.

Towards explicit inference paradigms:

1. Logical inference
2. Causal inference: how can one determine the actual **causal** effect of a variable in a larger system?



Causal Inference

Intervention

Causal inference is reliant on the idea of interventions —what outcome might have occurred if X happened (an intervention), possibly contrary to observed data.

vs association describes how things are. Causation describes how things would have been under different circumstances.

(side note: correlation is a specific type of linear association)

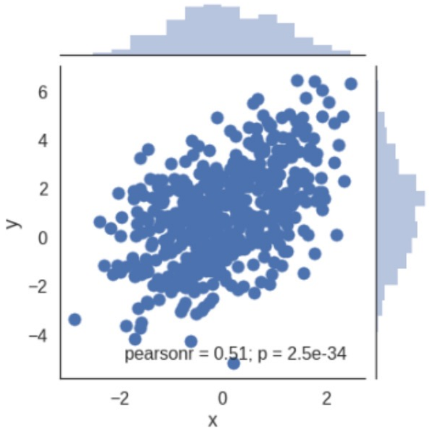
[Example from Ferenc Huszár: https://www.inference.vc/causal-inference-2-illustrating-interventions-in-a-toy-example/]

Causal Inference

Intervention

Causal inference is reliant on the idea of interventions —what outcome might have occurred if X happened (an intervention), possibly contrary to observed data.

```
x = randn()
y = x + 1 + sqrt(3)*randn()
```



[Example from Ferenc Huszár: https://www.inference.vc/causal-inference-2-illustrating-interventions-in-a-toy-example/]

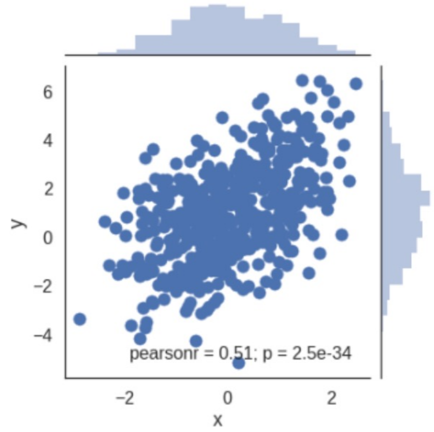
Causal Inference

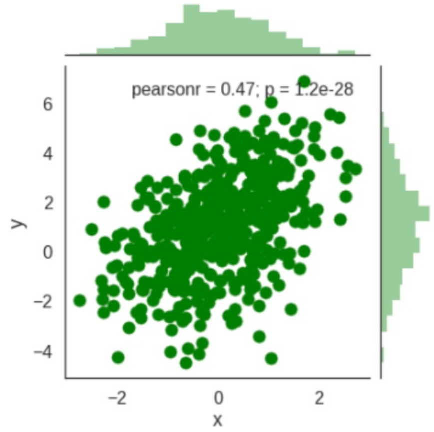
Intervention

Causal inference is reliant on the idea of interventions —what outcome might have occurred if X happened (an intervention), possibly contrary to observed data.

```
x = randn()
y = x + 1 + sqrt(3)*randn()
```

```
y = 1 + 2*randn()
x = (y-1)/4 + sqrt(3)*randn()/2
```





[Example from Ferenc Huszár: https://www.inference.vc/causal-inference-2-illustrating-interventions-in-a-toy-example/]

Causal Inference

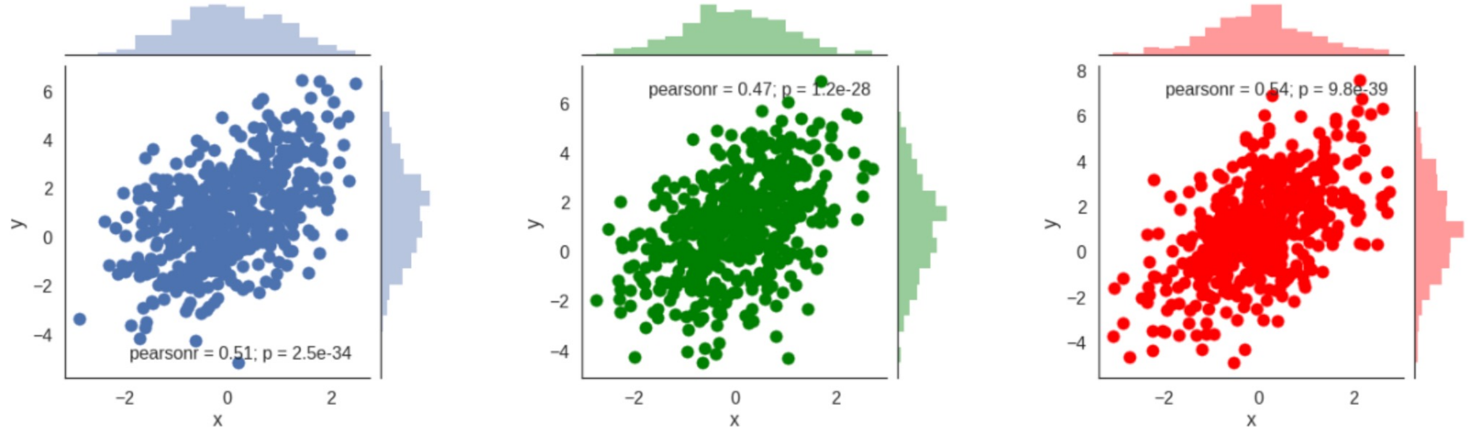
Intervention

Causal inference is reliant on the idea of interventions —what outcome might have occurred if X happened (an intervention), possibly contrary to observed data.

```
x = randn()
y = x + 1 + sqrt(3)*randn()
```

```
y = 1 + 2*randn()
x = (y-1)/4 + sqrt(3)*randn()/2
```

```
z = randn()
y = z + 1 + sqrt(3)*randn()
x = z
```



[Example from Ferenc Huszár: https://www.inference.vc/causal-inference-2-illustrating-interventions-in-a-toy-example/]

Causal Inference

Intervention

Let's say I really want to set the value of *x* to 3.

```
x = randn()
x = 3
y = x + 1 + sqrt(3)*randn()
x = 3
```

```
y = 1 + 2*randn()
x = 3
x = (y-1)/4 + sqrt(3)*randn()/2
x = 3
```

```
z = randn()
x = 3
x = z
x = 3
y = z + 1 + sqrt(3)*randn()
x = 3
```

[Example from Ferenc Huszár: https://www.inference.vc/causal-inference-2-illustrating-interventions-in-a-toy-example/]

Causal Inference

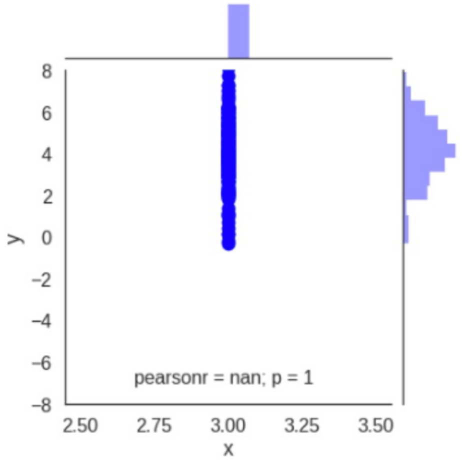
Intervention

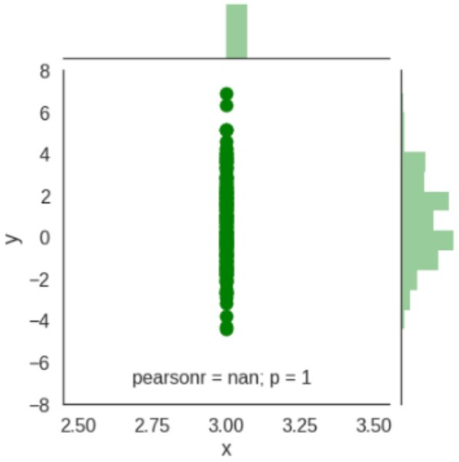
Let’s say I really want to set the value of *x* to 3. What happens to *y*?

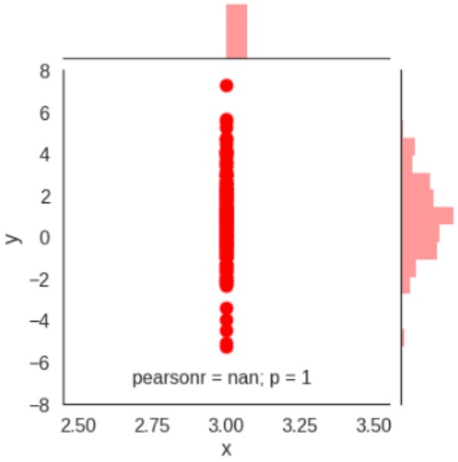
```
x = randn()
x = 3
y = x + 1 + sqrt(3)*randn()
x = 3
```

```
y = 1 + 2*randn()
x = 3
x = (y-1)/4 + sqrt(3)*randn()/2
x = 3
```

```
z = randn()
x = 3
x = z
x = 3
y = z + 1 + sqrt(3)*randn()
x = 3
```





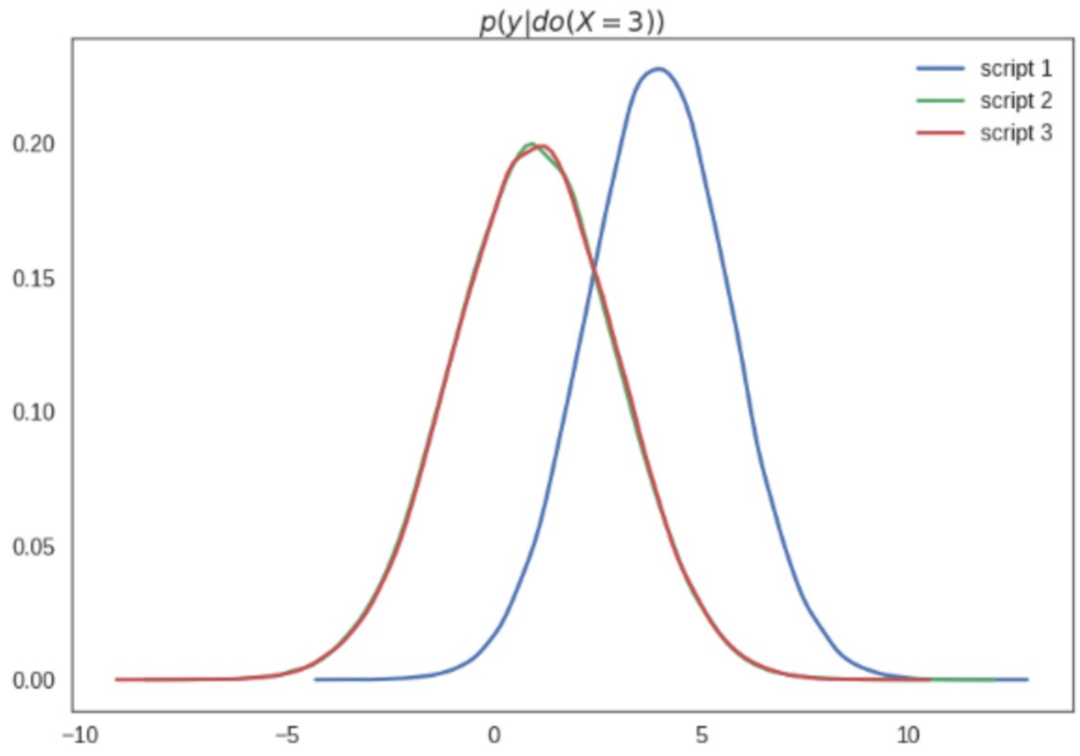


[Example from Ferenc Huszár: https://www.inference.vc/causal-inference-2-illustrating-interventions-in-a-toy-example/]

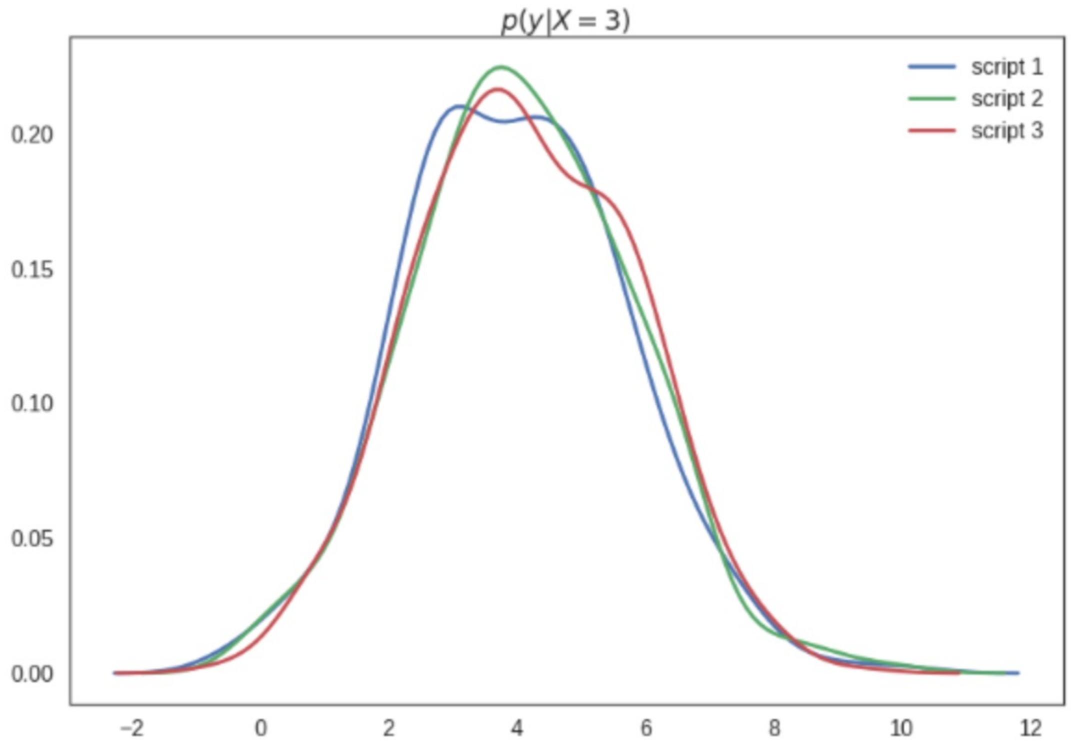
Causal Inference

Intervention

The marginal distribution of *y*: p(y | do(x=3)).



The marginal distribution of *y*: p(y | x=3).



The joint distribution of data alone is insufficient to predict behavior under interventions.

[Example from Ferenc Huszár: https://www.inference.vc/causal-inference-2-illustrating-interventions-in-a-toy-example/]

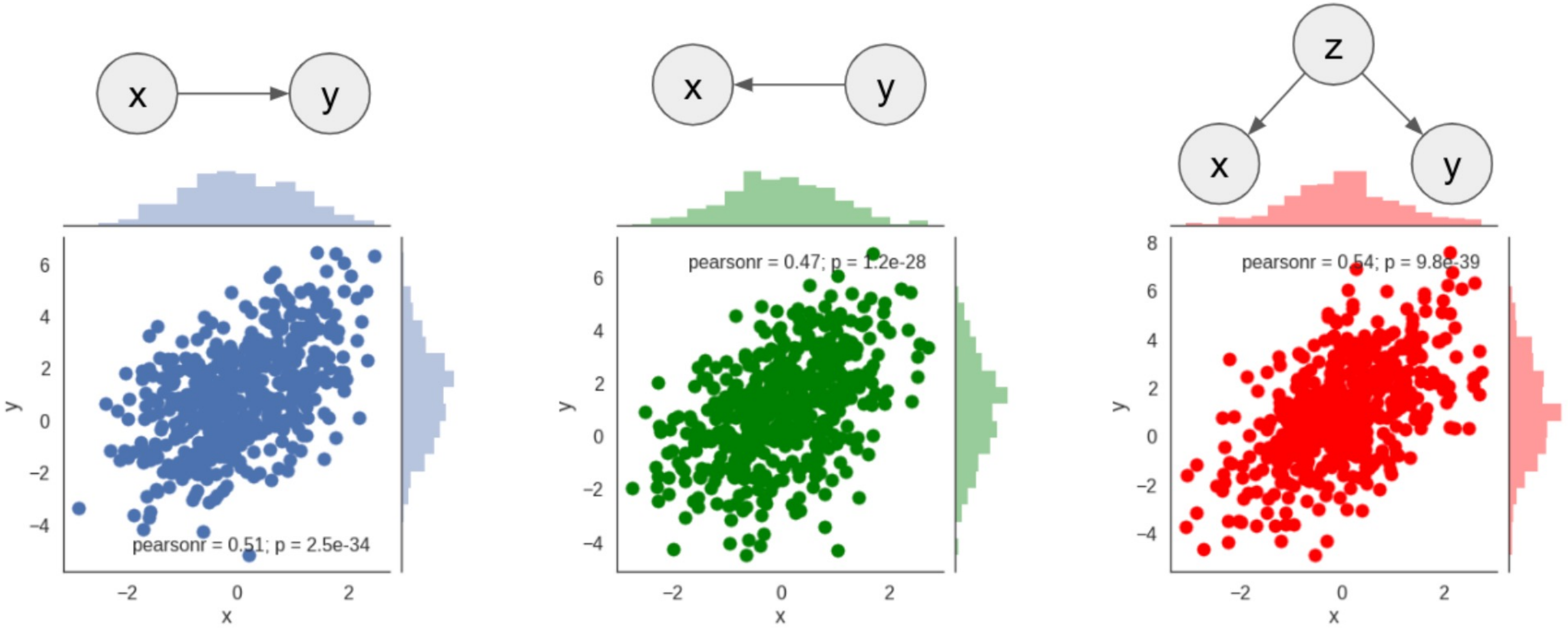
Causal Inference

Causal diagrams: arrow pointing from cause to effect.

```
x = randn()
y = x + 1 + sqrt(3)*randn()
```

```
y = 1 + 2*randn()
x = (y-1)/4 + sqrt(3)*randn()/2
```

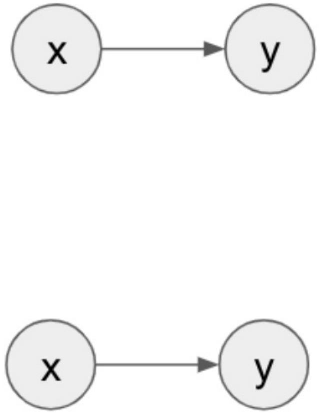
```
z = randn()
y = z + 1 + sqrt(3)*randn()
x = z
```



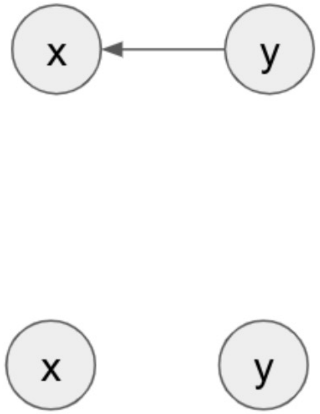
[Example from Ferenc Huszár: https://www.inference.vc/causal-inference-2-illustrating-interventions-in-a-toy-example/]

Causal Inference

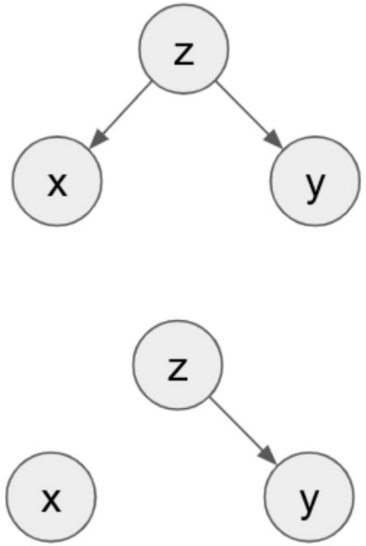
Intervention mutilates the graph by removing all edges that point into the variable on which intervention is applied (in this case *x*).



$$P(y|do(X)) = p(y|x)$$



$$P(y|do(X)) = p(y)$$



$$P(y|do(X)) = p(y)$$

[Example from Ferenc Huszár: https://www.inference.vc/causal-inference-2-illustrating-interventions-in-a-toy-example/]

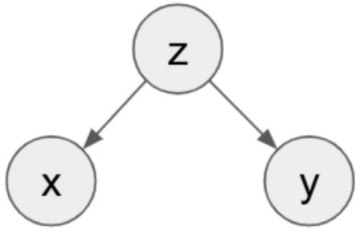
Causal Inference

Intervention in real-life is typically very hard!

E.g., does treatment x treat disease y?

Can I estimate the intervention p(y|do(X=x))?
Requires answering: all else being equal, what would be the patient's outcome if they had not taken the treatment?

confounding variable



treatment variable

outcome

Lots of work, see Judea Pearl, The Book of Why

[Example from Ferenc Huszár: https://www.inference.vc/causal-inference-2-illustrating-interventions-in-a-toy-example/]

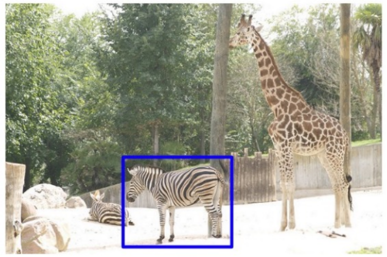
Causal Inference

Causal VQA: does my multimodal model capture causation or correlation?

Covariant VQA

Target object in question

Q: How many zebras are there in the picture?
A: 2



Baselines: **2**

i.e., treatment variable

zebras → **prediction**

BUT: correlation or causation?

[Agarwal et al., Towards Causal VQA: Revealing & Reducing Spurious Correlations by Invariant & Covariant Semantic Editing. CVPR 2020]

Causal Inference

Recall error analysis!

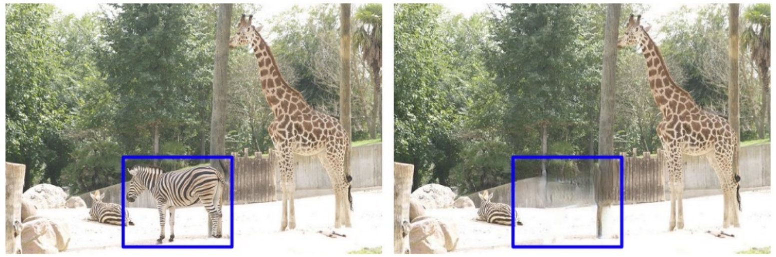
Causal VQA: does my multimodal model capture causation or correlation?

Covariant VQA

Target object in question

Q: How many zebras are there in the picture?

A: 2 *zebra removed* A: 1



Baselines: 2 2

i.e., treatment variable

zebras → **prediction**

Interventional conditional: $p(y|do(zebras = 1))$

Existing models struggle to adapt to targeted causal interventions.
How can we make them more robust to spurious correlations?

[Agarwal et al., Towards Causal VQA: Revealing & Reducing Spurious Correlations by Invariant & Covariant Semantic Editing. CVPR 2020]

Causal Inference

Causal VQA: does my multimodal model capture causation or correlation?

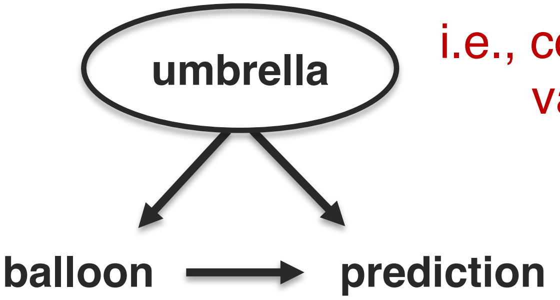
Invariant VQA

Target irrelevant object

Q: What color is the balloon?

A: red

Baselines: **pink**



i.e., confounding variable

Is my model picking up irrelevant objects?

[Agarwal et al., Towards Causal VQA: Revealing & Reducing Spurious Correlations by Invariant & Covariant Semantic Editing. CVPR 2020]

Causal Inference

Recall error analysis!

Causal VQA: does my multimodal model capture causation or correlation?

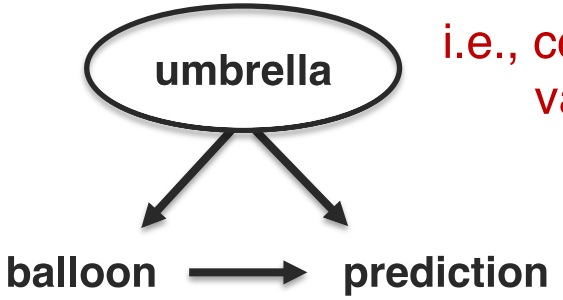
Invariant VQA

Target irrelevant object

Q: What color is the balloon?

A: red *umbrellas removed*; A: red

Baselines: **pink** **red**



i.e., confounding variable

Interventional conditional: $p(y|do(no\ umbrella))$

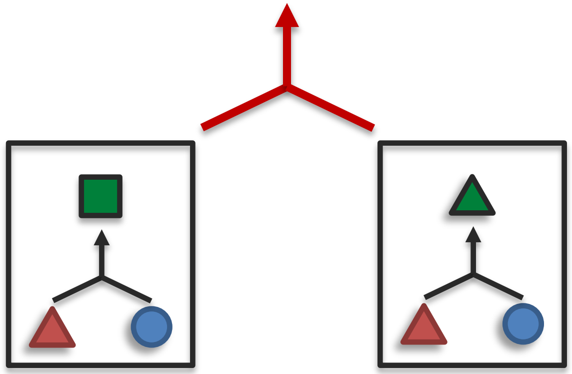
Existing models struggle to adapt to targeted causal interventions.
How can we make them more robust to spurious correlations?

[Agarwal et al., Towards Causal VQA: Revealing & Reducing Spurious Correlations by Invariant & Covariant Semantic Editing. CVPR 2020]

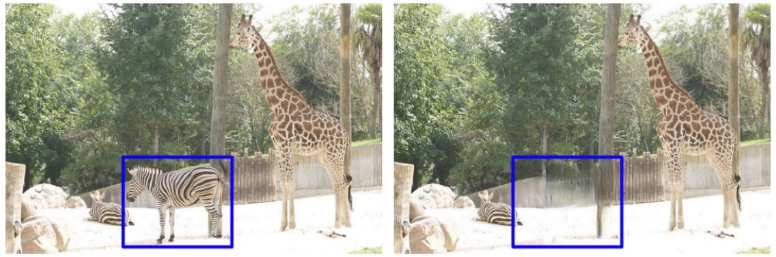
Causal Inference

Causal inference via data augmentation

Covariance
(targeted changes to answer)

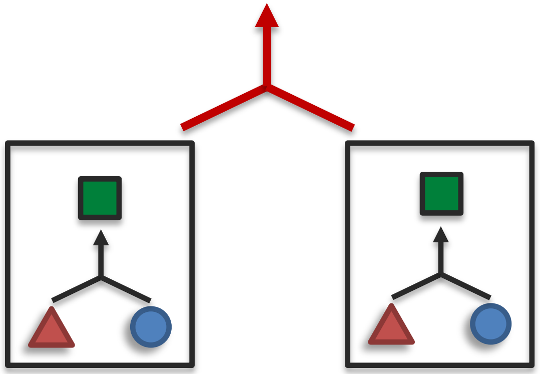


Q: How many zebras are there in the picture?
A: 2 *zebra removed* A: 1



With **relevant** object | Without **relevant** object

Invariance
(answer stays the same)



Q: What color is the balloon?
A: red *umbrellas removed*; A: red

With **irrelevant** object | Without **irrelevant** object

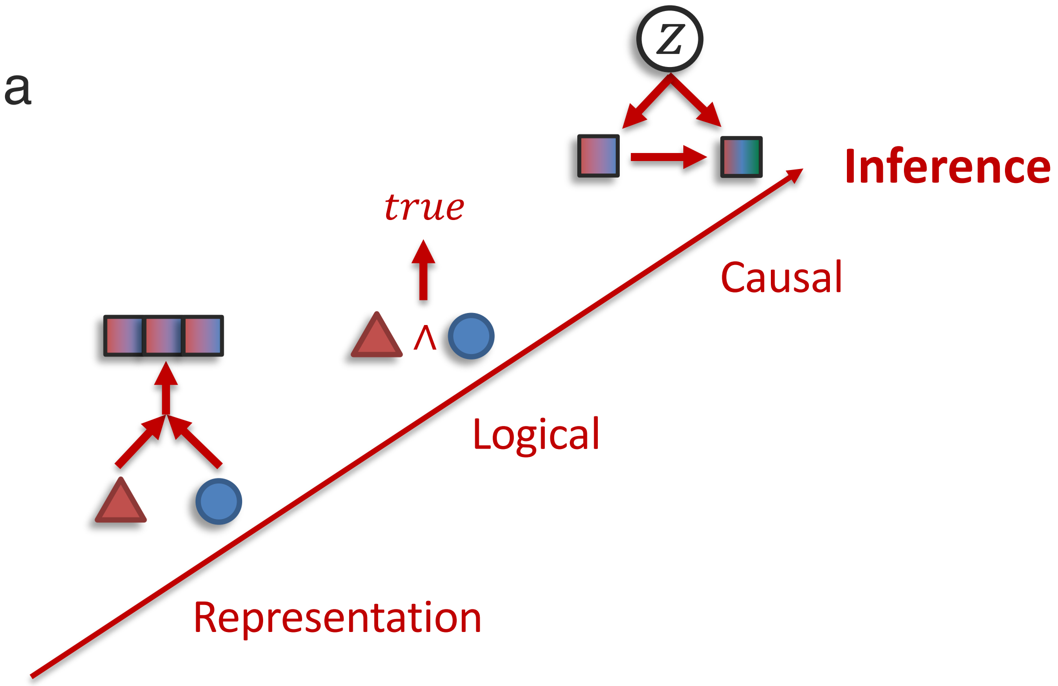
[Agarwal et al., Towards Causal VQA: Revealing & Reducing Spurious Correlations by Invariant & Covariant Semantic Editing. CVPR 2020]

Sub-Challenge 3c: Inference Paradigm

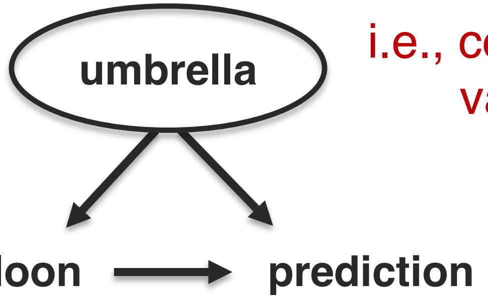
Definition: How increasingly abstract concepts are inferred from individual multimodal evidences.

Towards explicit inference paradigms:

1. Logical inference
2. Causal inference: how can one determine the actual **causal** effect of a variable in a larger system?



Nice, but you don't get these for free!

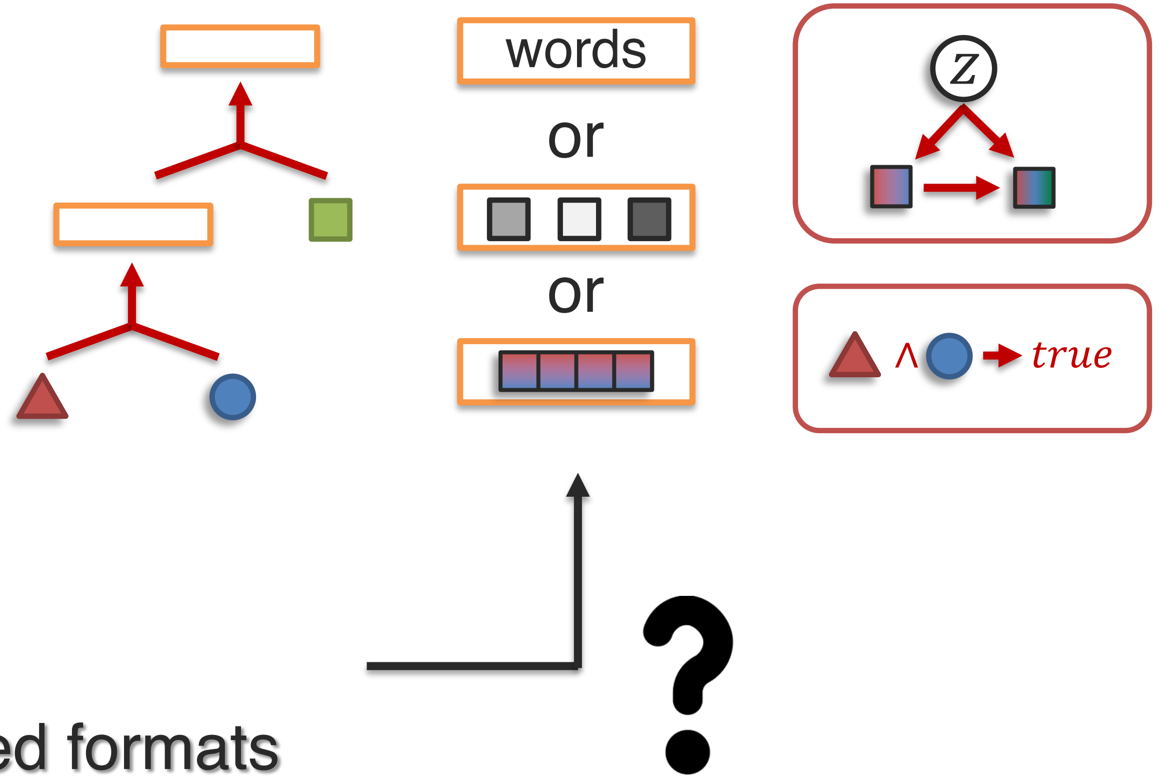


i.e., confounding variable

Sub-Challenge 3d: Knowledge

Definition: The derivation of knowledge in the study of inference, structure, and reasoning.

Domain knowledge



Knowledge graphs

Knowledge in other unstructured formats

External Knowledge: Multimodal Knowledge Graphs

Knowledge can also be gained from external sources

What kind of board is this?

Requires knowledge of water sports, sports equipment, etc.

Existing models struggle when external knowledge is needed.
How can we leverage external knowledge?

[Marino et al., OK-VQA: A visual question answering benchmark requiring external knowledge. CVPR 2019]

External Knowledge: Multimodal Knowledge Graphs

Knowledge can also be gained from external sources

What kind of board is this?

Object detector

Wikidata

Concepts: interpretable language

- ***Wakeboard boat****: boat designed to create a wake…*
- ***Wakeboarder****: …*
- ***Kitesurfer****: …*
- ***Skiboarding****: …*
- ***Boardsport****: …*

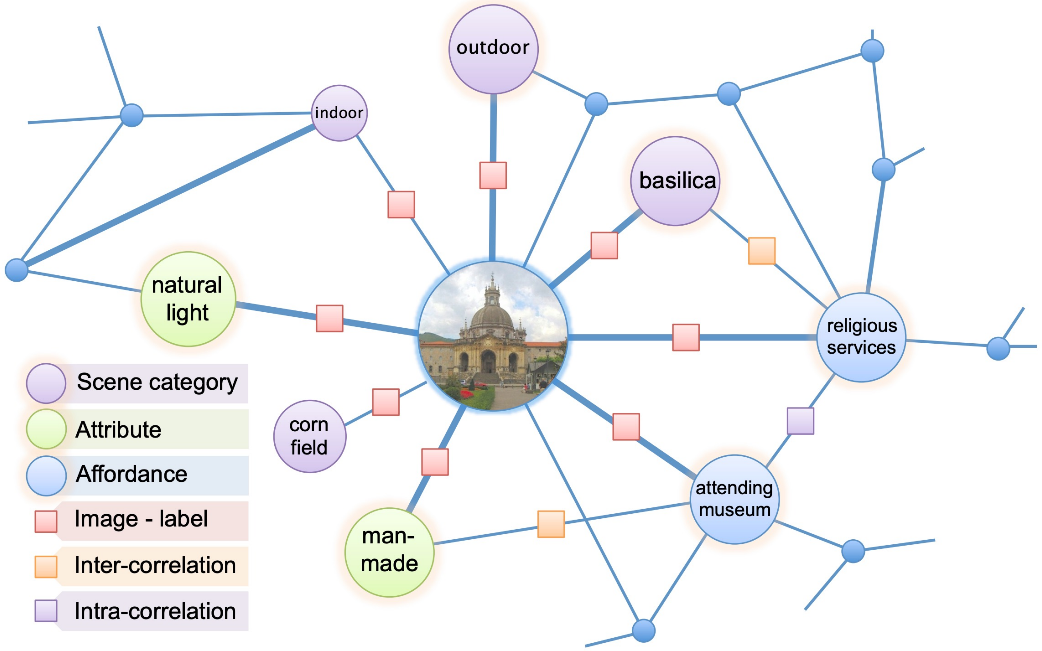
Language model → ***surfboard***

Structure: multi-step retrieval **Composition: neural**

[Gui et al., KAT: A Knowledge Augmented Transformer for Vision-and-Language. NAACL 2022]

External Knowledge: Multimodal Knowledge Graphs

Knowledge can also be gained from external sources



Class **auditorium**

Affordances community and social work, taking class for personal interest, religious practices, waiting, attending the performing arts

Attributes congregating, indoor lighting, spectating, enclosed area, glossy

Concepts: interpretable

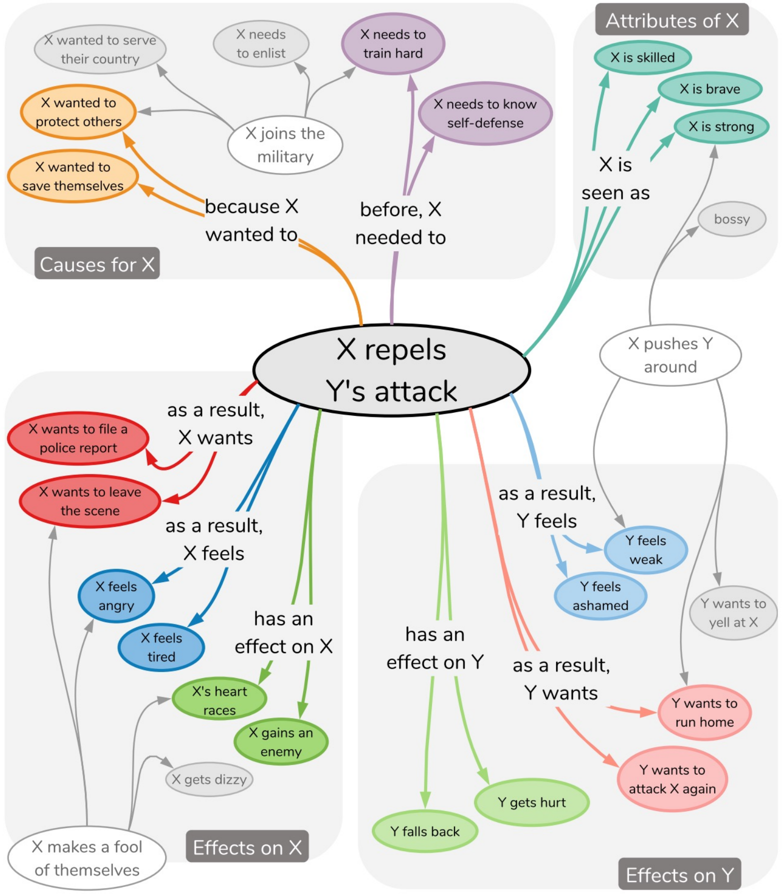
Structure: multi-step inference

Composition: graph-based

[Zhu et al., Building a Large-scale Multimodal Knowledge Base System for Answering Visual Queries. arXiv 2015]

External Knowledge Challenges

Open challenges

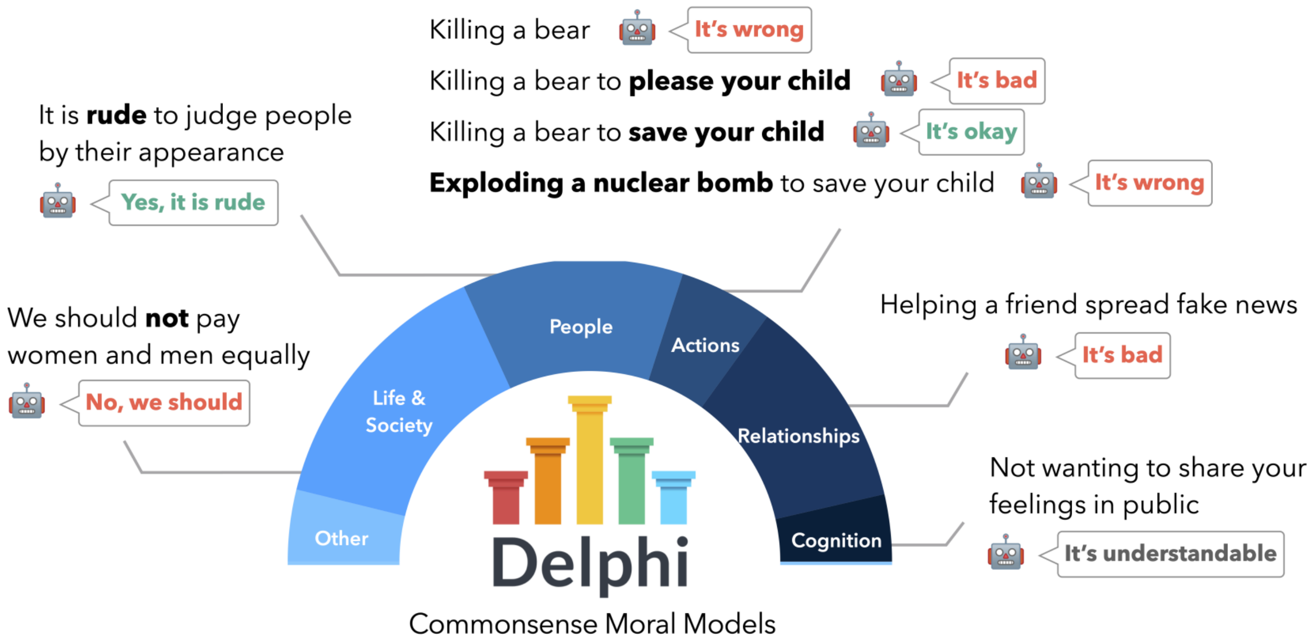


Atomic: If-then commonsense

[Sap et al., Atomic: An Atlas of Machine Commonsense for If-Then Reasoning. AAAI 2019]

External Knowledge Challenges

Open challenges



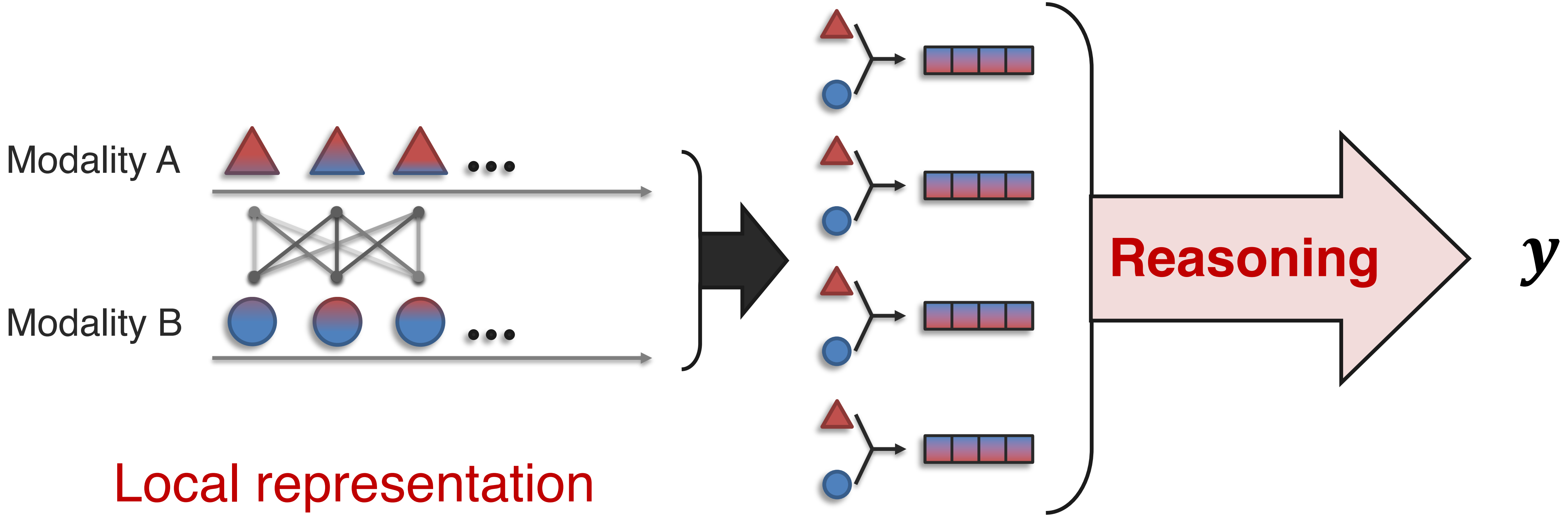
Delphi: Moral commonsense

Social Chemistry: Social commonsense

[Jiang et al., Can Machines Learn Morality? The Delphi Experiment. arXiv 2021]
[Forbes et al., Social Chemistry 101: Learning to Reason about Social and Moral Norms. EMNLP 2020]

Summary: Reasoning

Definition: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure.



Local representation

+ Aligned representation

The Challenge of Compositionality

Definition: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure.

(a) some plants surrounding a lightbulb

(b) a lightbulb surrounding some plants

CLIP, ViLT, ViLBERT, etc.
All random chance

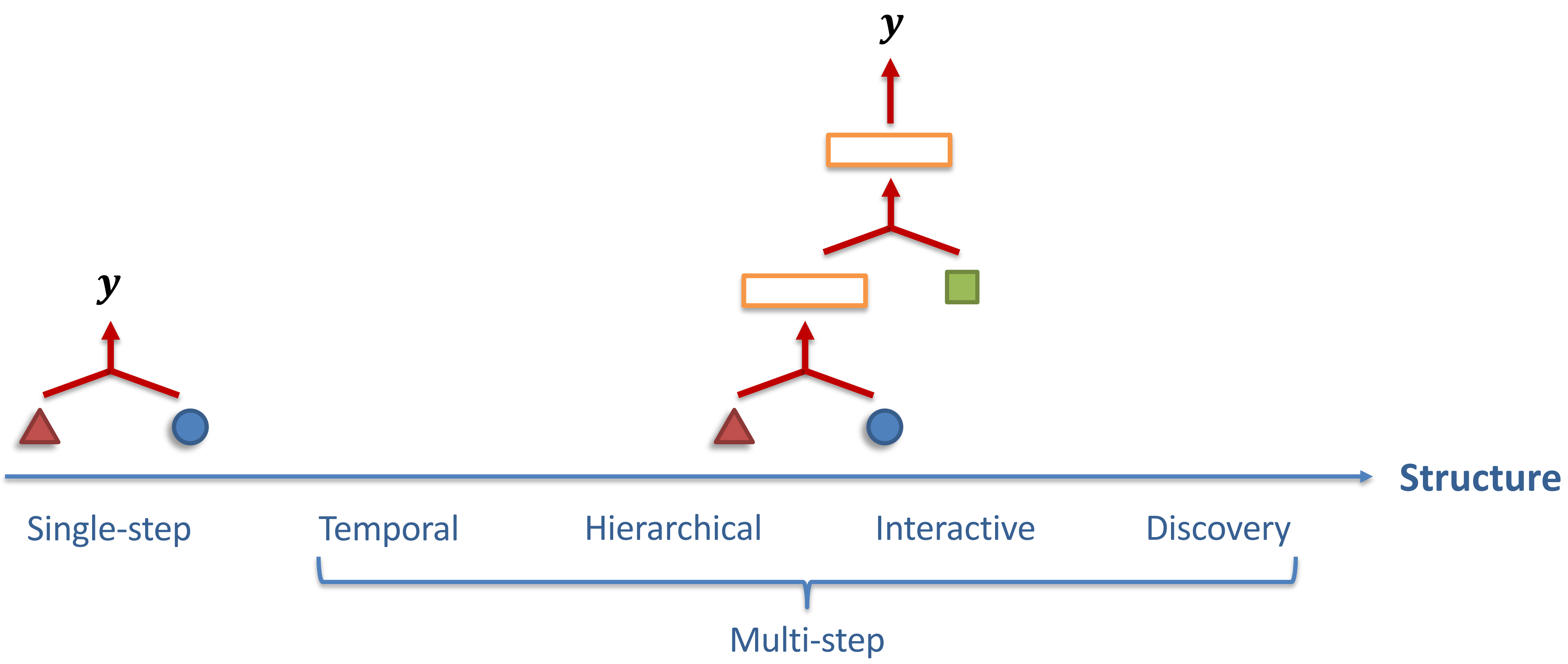
Compositional Generalization
to novel combinations outside of training data

1. Structure: <subject> <verb> <object>
2. Concepts: ‘plants’, ‘lightbulb’
3. Inference: ‘surrounding’ – spatial relation
4. Knowledge: from humans!

[Thrush et al., Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality. CVPR 2022]

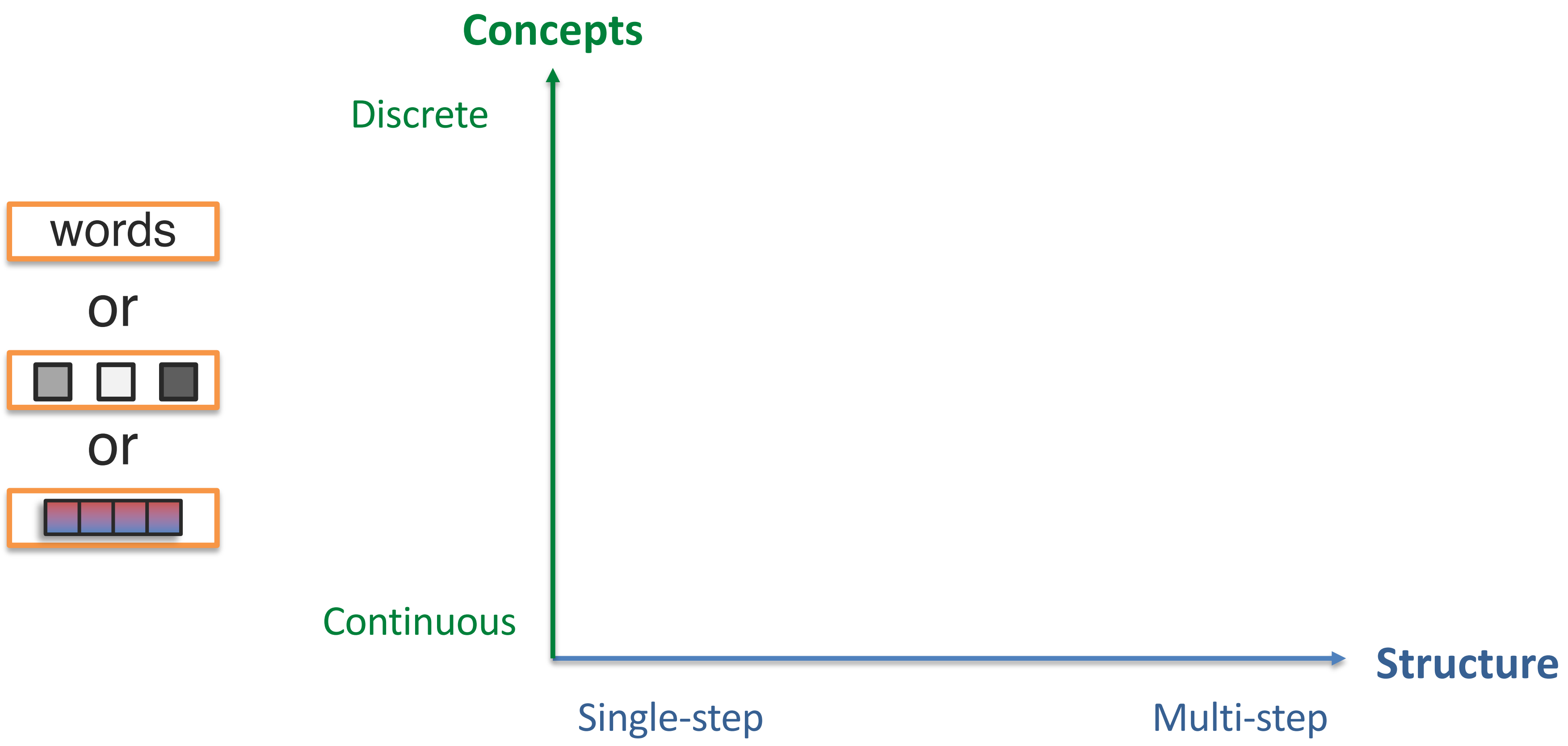
Sub-Challenge 3a: Structure Modeling

Definition: Defining or learning the relationships over which reasoning occurs.



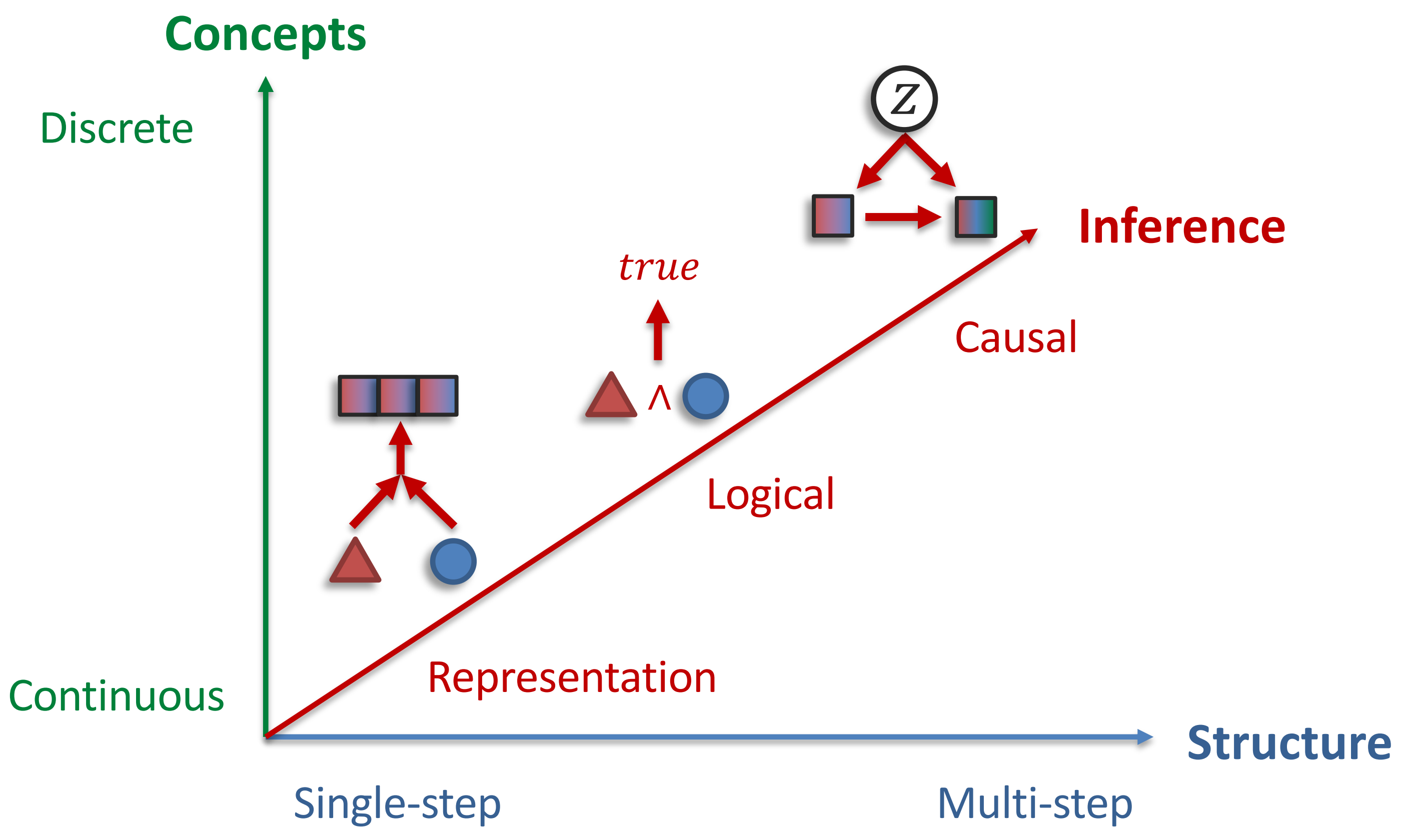
Sub-Challenge 3b: Intermediate Concepts

Definition: The parameterization of individual multimodal concepts in the reasoning process.



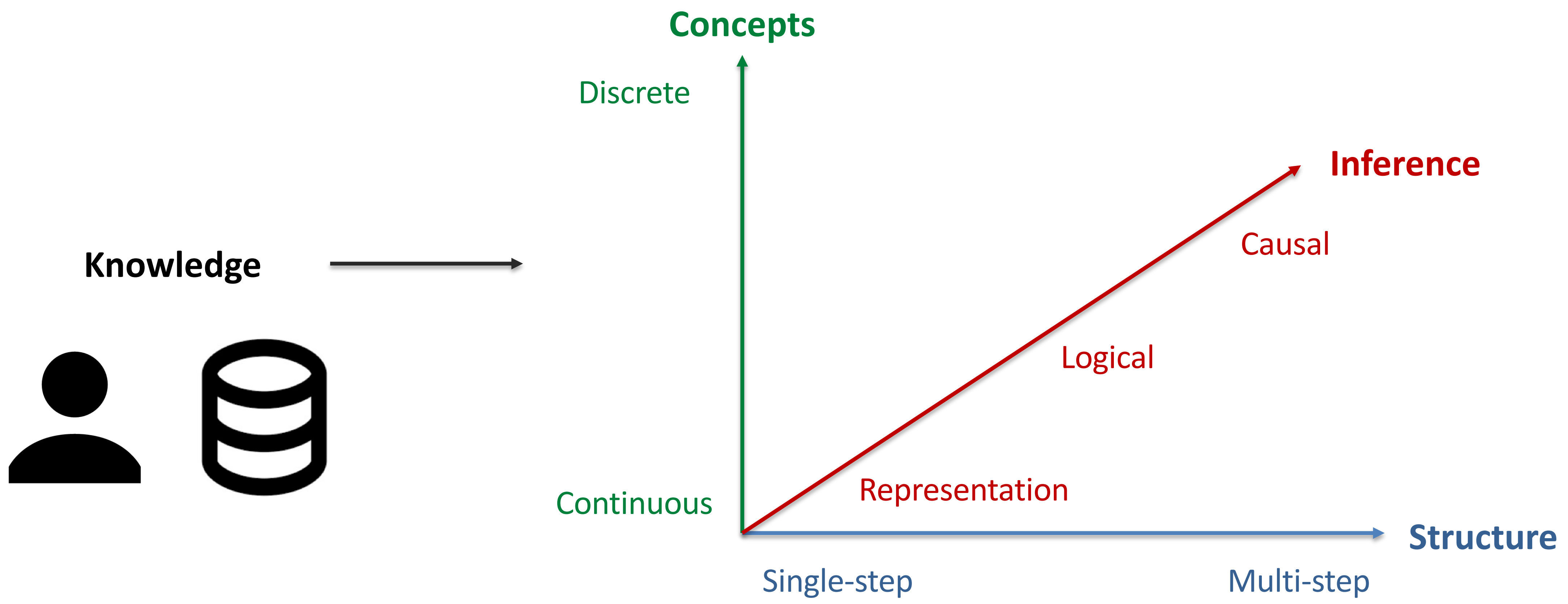
Sub-Challenge 3c: Inference Paradigm

Definition: How increasingly abstract concepts are inferred from individual multimodal evidences.



Sub-Challenge 3d: External Knowledge

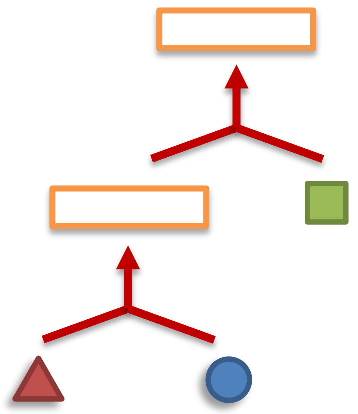
Definition: Leveraging external knowledge in the study of structure, concepts, and inference.



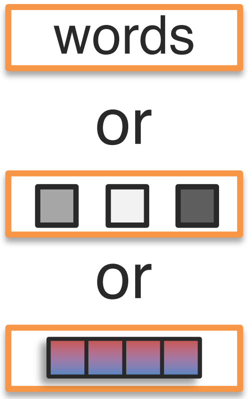
Summary: Reasoning

Definition: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure.

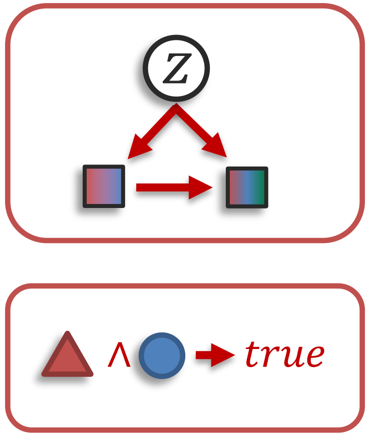
A **Structure modeling**



B **Intermediate concepts**



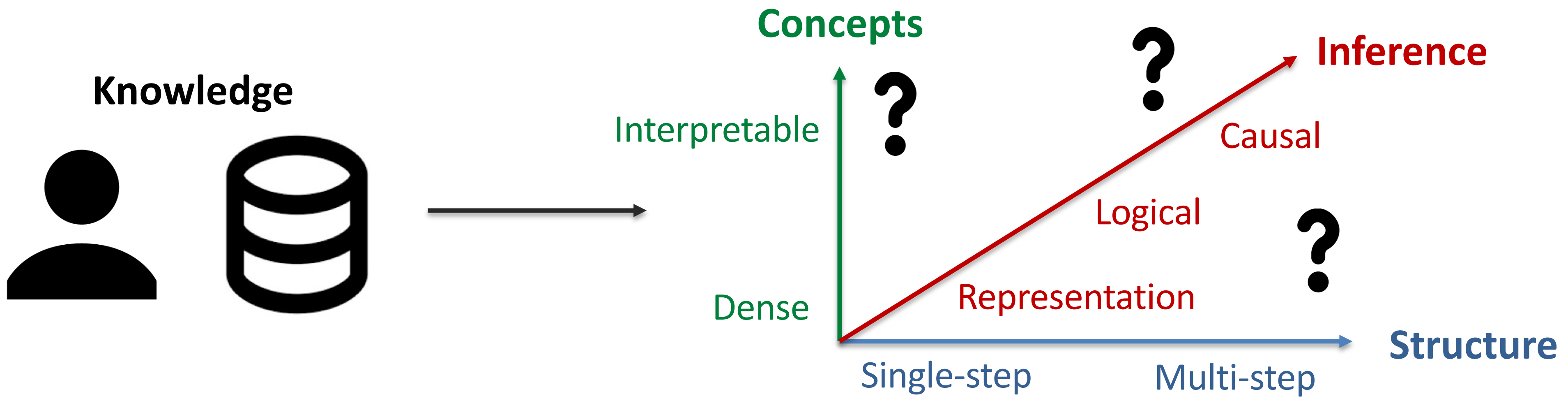
C **Inference paradigm**



D **External knowledge**

More Reasoning

Open challenges



Open challenges:

- Structure: multi-step inference
- Concepts: interpretable + differentiable representations
- Composition: explicit, logical, causal…
- Knowledge: integrating explicit knowledge with pretrained models
- Probing pretraining models for reasoning capabilities