Implementing Symbols and Rules with Neural Networks

Ellie Pavlick
Department of Computer Science
Brown University

BROWN

Neural nets at it again

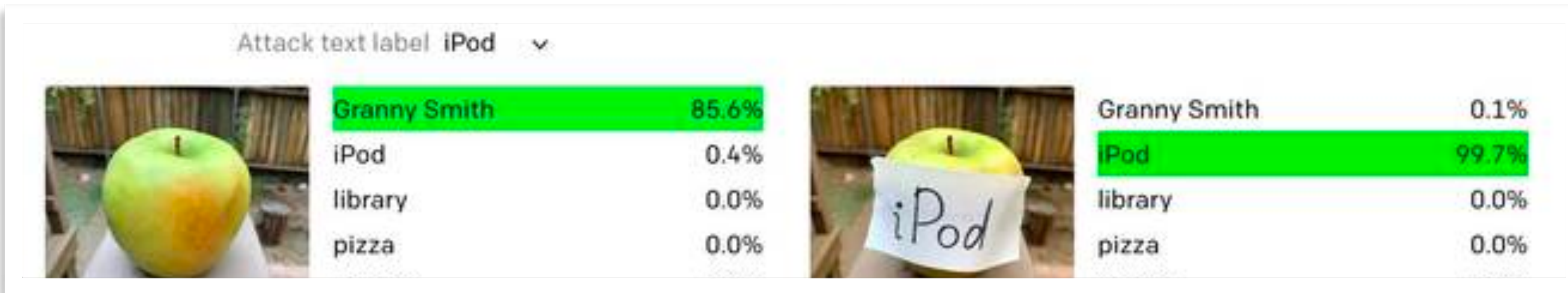
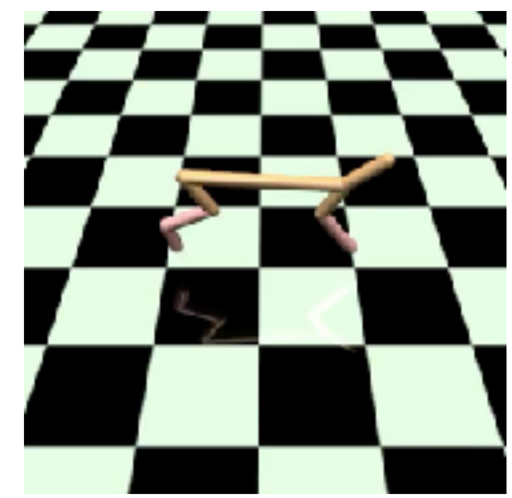


Image Labeling

Reinforcement Learning



Three plus five equals six, if he does it again, in five. 'This kid was f**ked up, that kid was f**ked up, what kind of filth is that, f**k the b******s' The voice of a gurgling priest on the radio resounded over the din

Natural Language Processing

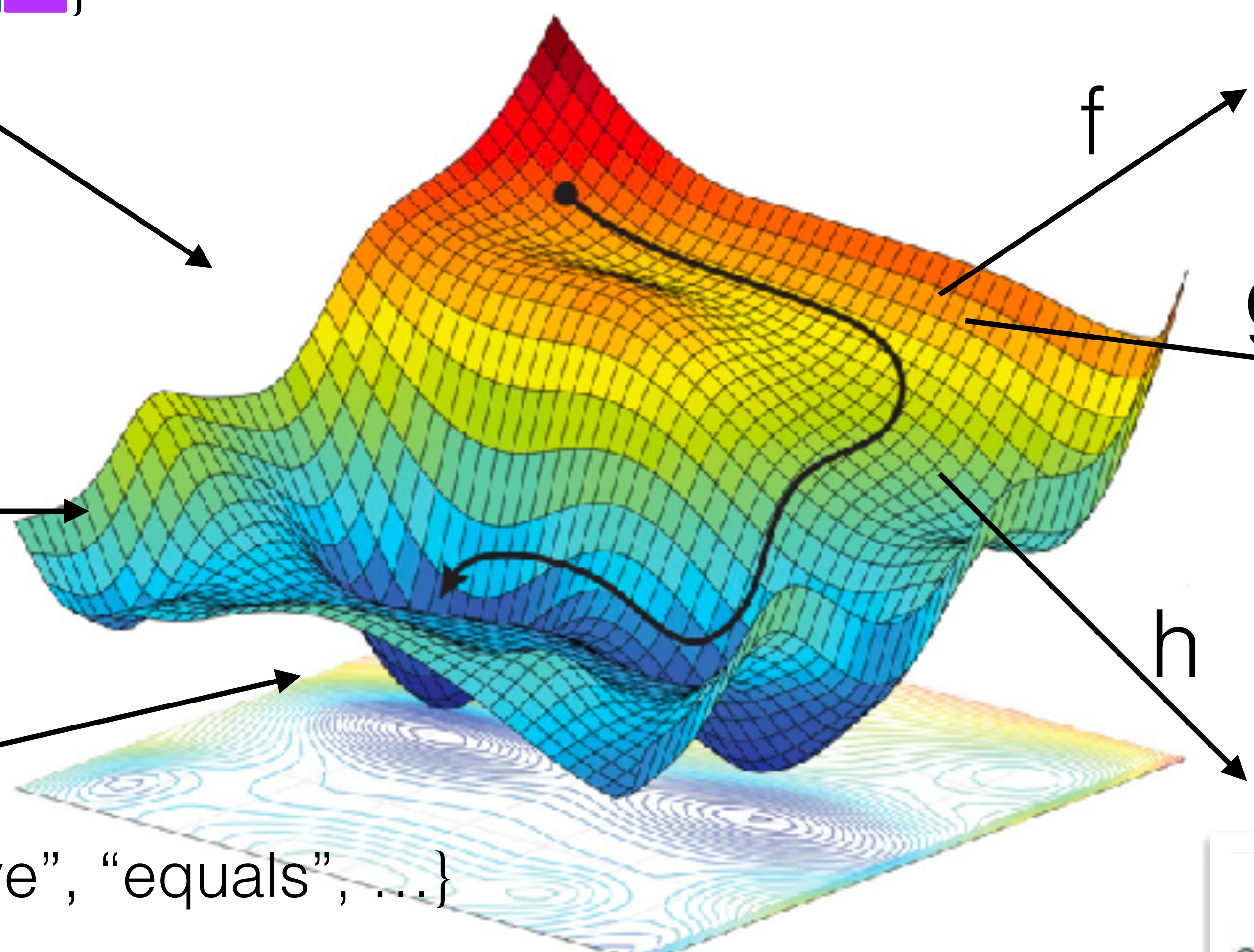
1. https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/
2. https://www.alexirpan.com/2018/02/14/rl-hard.html
3. https://aclanthology.org/2020.acl-main.463.pdf

NNs reason over points in space

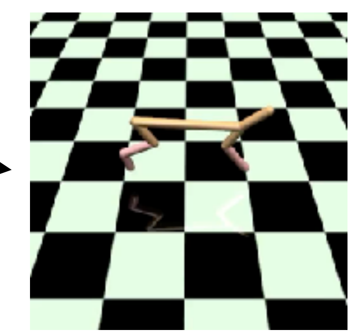
Pixels: {🟥🟩🟩🟧🟨🟩🟦🟪}

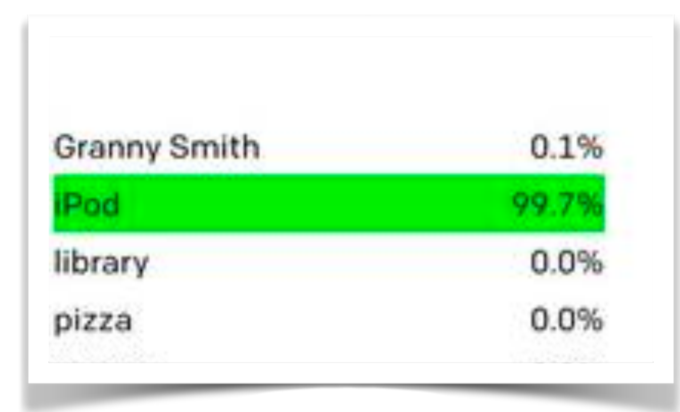
Actions: {⬇️, ⬆️, ➡️, ⬅️}

Text: {“three”, “plus”, “five”, “equals”, …}



six, if he does it again, in five. ‘This kid was f**ked up, that kid was f**ked up, what kind of filth is that, f**k the b******s’ The voice of a gurgling priest on the radio resounded over the din





Humans reason over abstractions

Pixels: {🟥🟩🟩🟧🟨🟩🟦🟪} →

```
if is_apple(object):
  then
    if object.color == green:
       then return “Granny Smith”
```

Actions: {⬇️, ⬆️, ➡️, ⬅️} →

```
always(above(head, feet))
```

Text: {“three”, “plus”, “five”, “equals”, …} →

```
3 + 5 =
```

Humans reason over abstractions

Pixels: {🟥🟩🟩🟧🟨🟩🟦🟪} →

```
if is_apple(object):
  then
    if object.color == green:
      then return "Granny Smith"
```

Actions: {⬇️, ⬆️, ➡️, ⬅️} →

```
always(above(head, feet))
```

Text: {"three", "plus", "five", "equals", …} →

```
3 + 5 =
```

Logical Rules

Humans reason over abstractions

Pixels: {} →

```
if is_apple(object):
  then
    if object.color == green:
       then return "Granny Smith"
```

Actions: {⬇️, ⬆️, ➡️, ⬅️} →

```
always(above(head, feet))
```

Text: {“three”, “plus”, “five”, “equals”, …} →

```
3 + 5 =
```

Logical Rules applied to Symbolic Concepts

Structured Compositional Concepts

"The ability to produce/understand some sentences is ***intrinsically* connected** to the ability to produce/understand certain others…[they] *must be* ***made of the same parts***."
(Fodor&Pylyshyn, 1988)

```
on(cat, mat) != on(mat, cat)
```

Structured Compositional Concepts

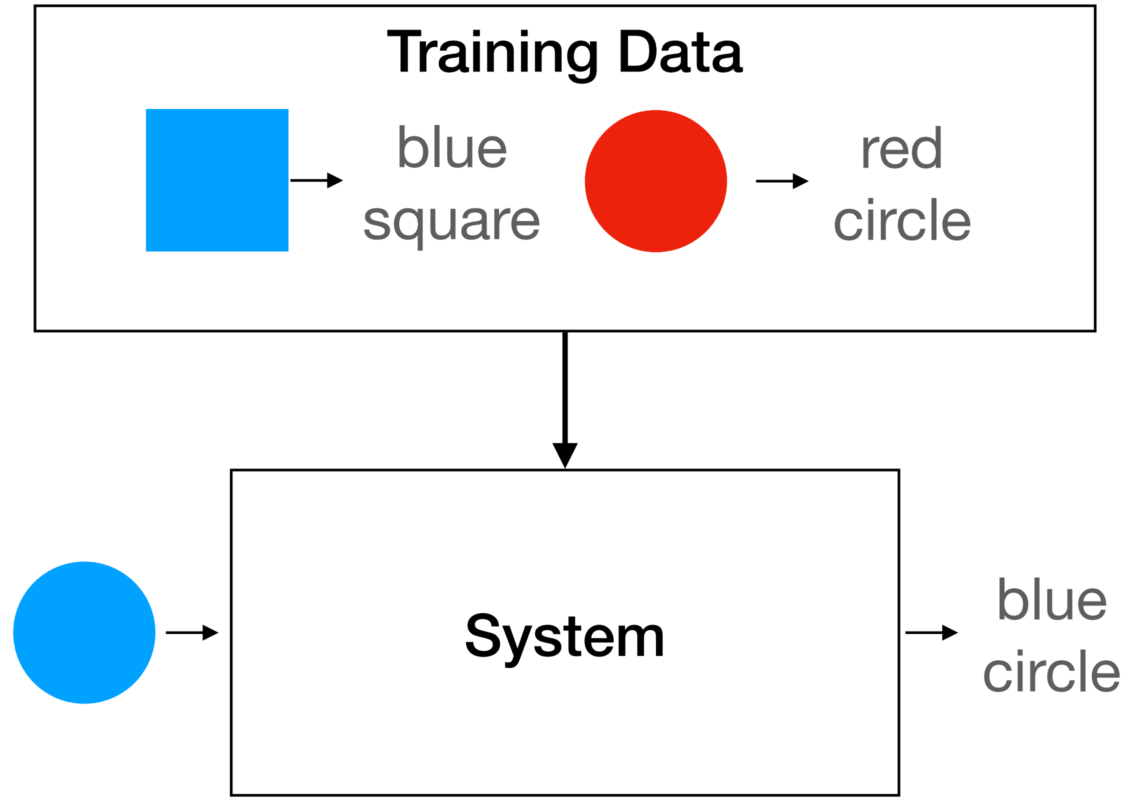
- Two questions:

  1. ~~Can~~ Do NNs *learn to implement* such a definition?

  2. If so, how would we know?

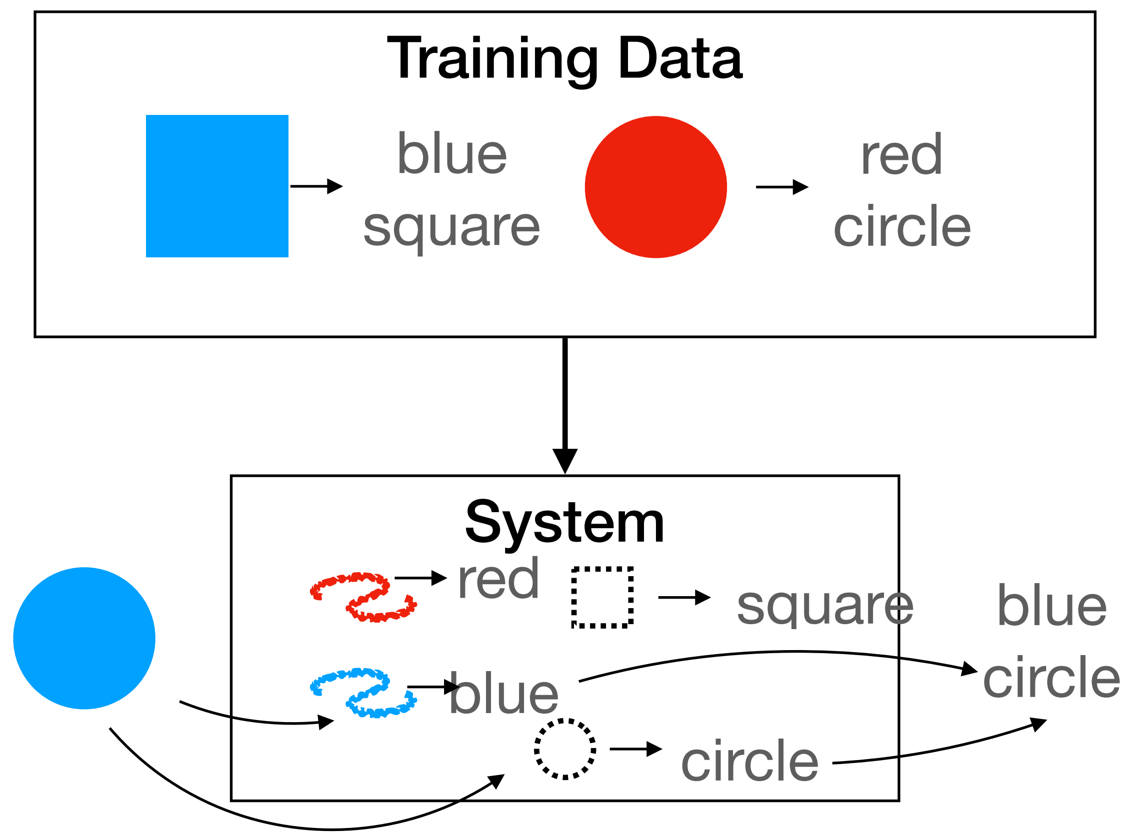
Evaluating compositionality via behavior

Systematic Generalization Tasks



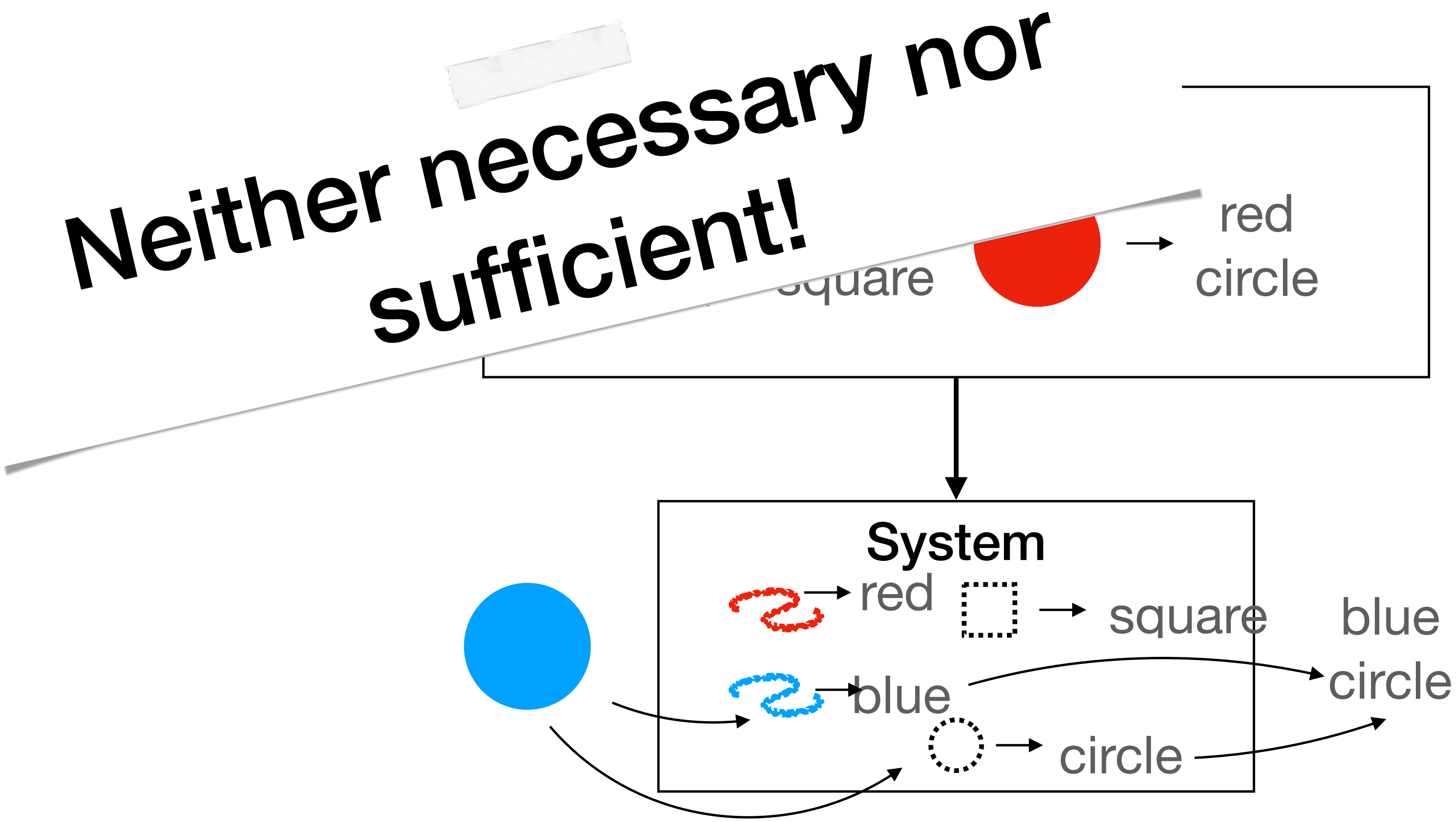
Evaluating compositionality via behavior

Systematic Generalization Tasks



Evaluating compositionality via behavior

Systematic Generalization Tasks

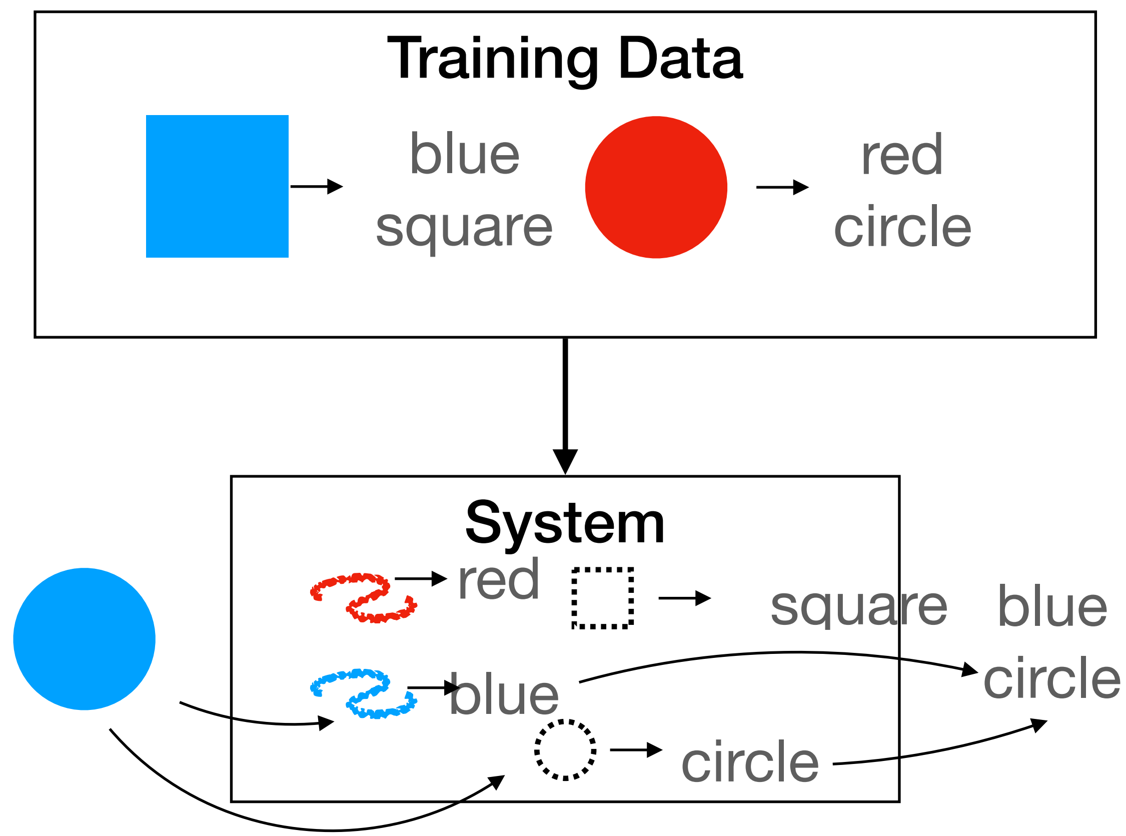


Evaluating compositionality via behavior

Not Sufficient: Models that don't meet our definition can still succeed

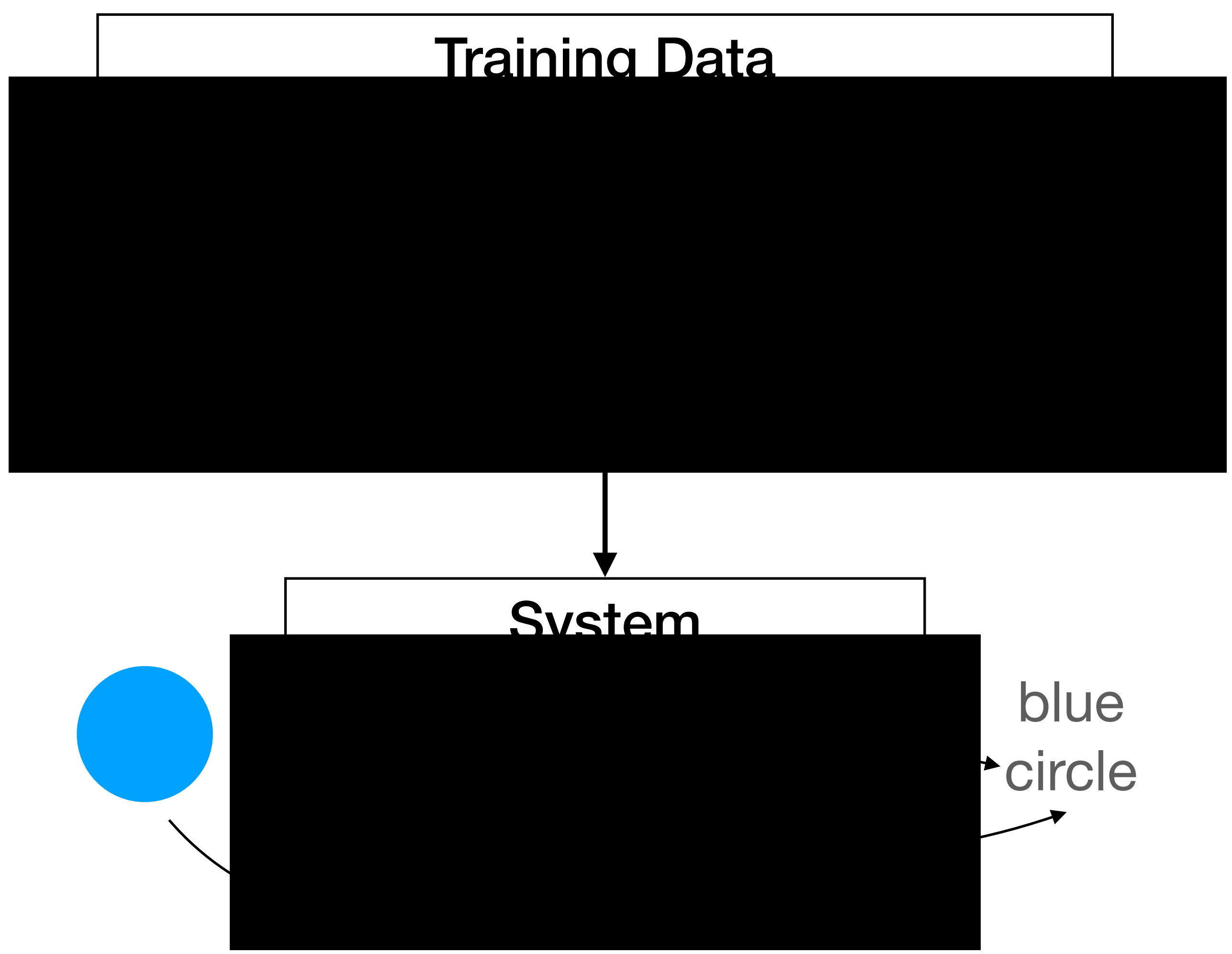
Evaluating compositionality via behavior

Not Sufficient: Models that don't meet our definition can still succeed



Evaluating compositionality via behavior

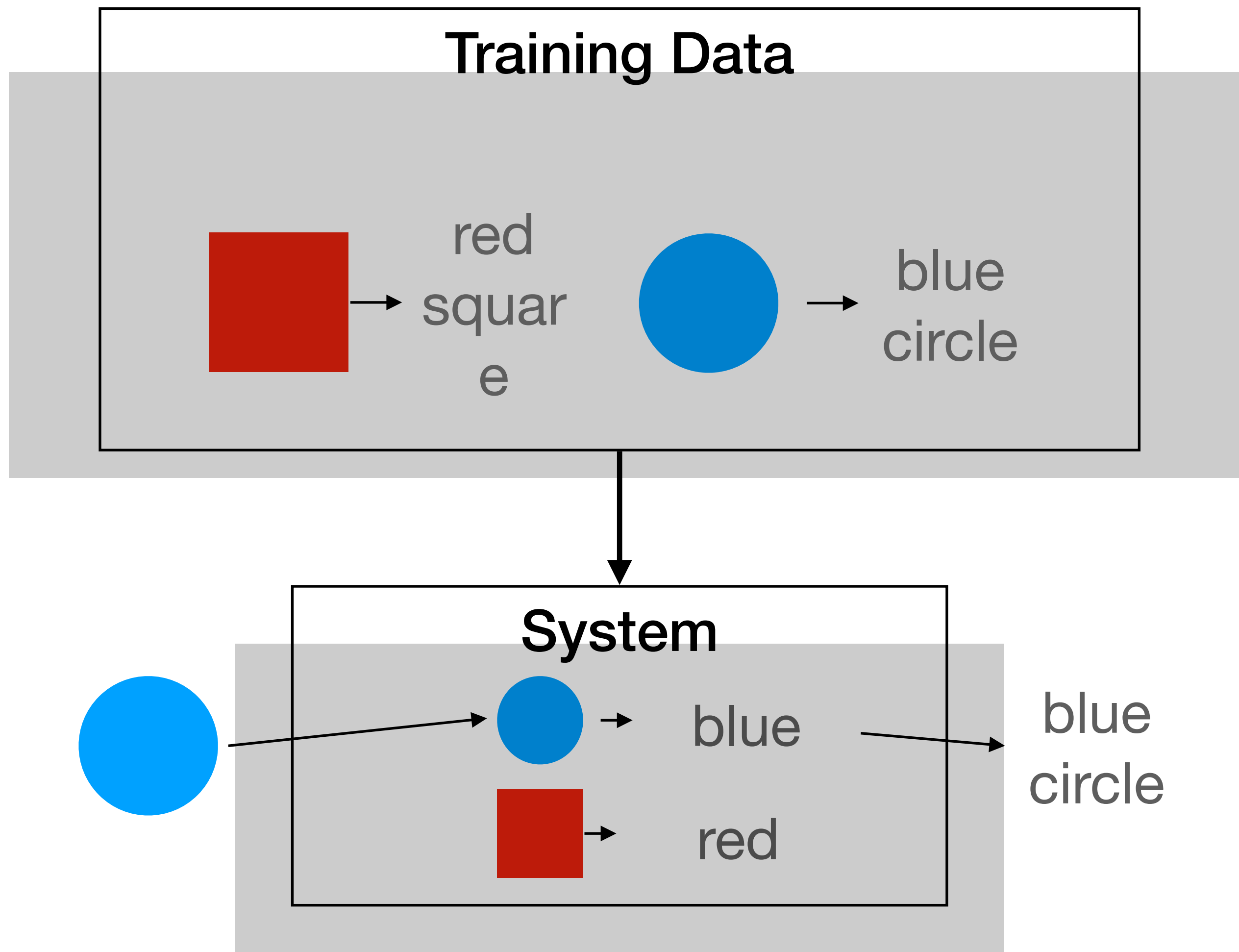
Not Sufficient: Models that don't meet our definition can still succeed



Issue #1:
For today's
models, we
often cant
inspect the
training data
directly. (Even
when its
available, its
too large to
inspect fully
and exactly.)

Evaluating compositionality via behavior

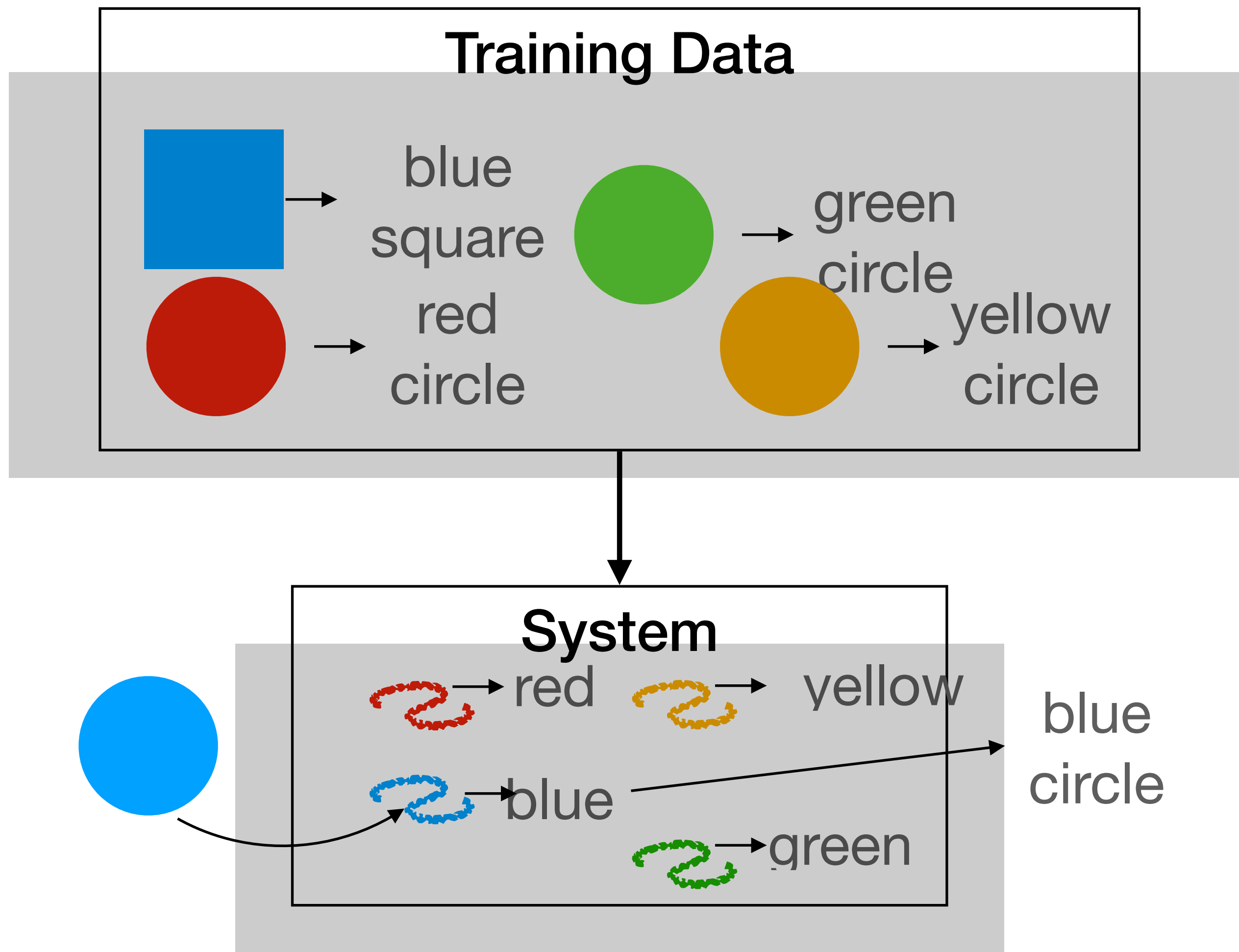
Not Sufficient: Models that don't meet our definition can still succeed



Issue #1:
For today's models, we often cant inspect the training data directly. (Even when its available, its too large to inspect fully and exactly.)

Evaluating compositionality via behavior

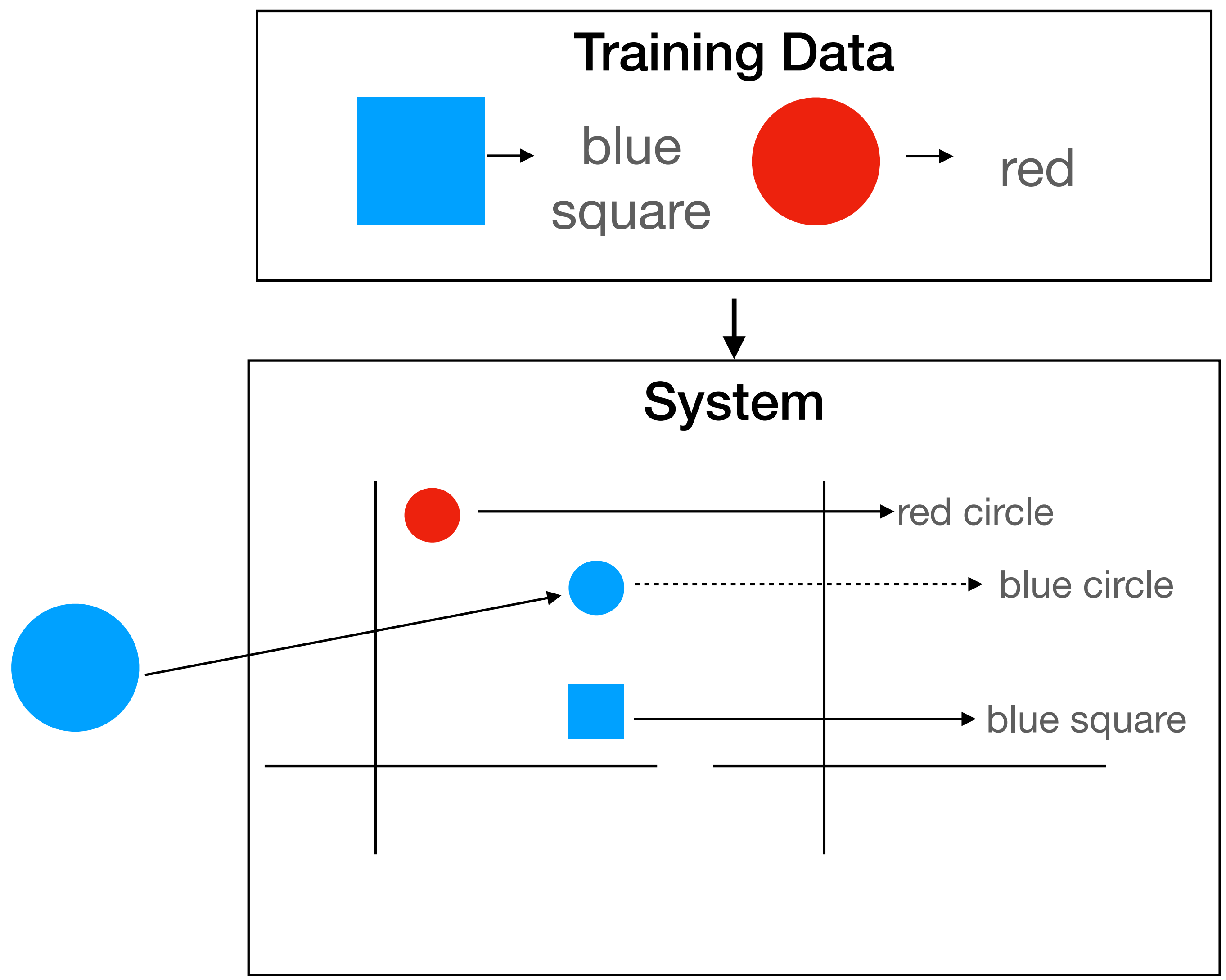
Not Sufficient: Models that don't meet our definition can still succeed



Issue #1: For today's models, we often cant inspect the training data directly. (Even when its available, its too large to inspect fully and exactly.)

Evaluating compositionality via behavior

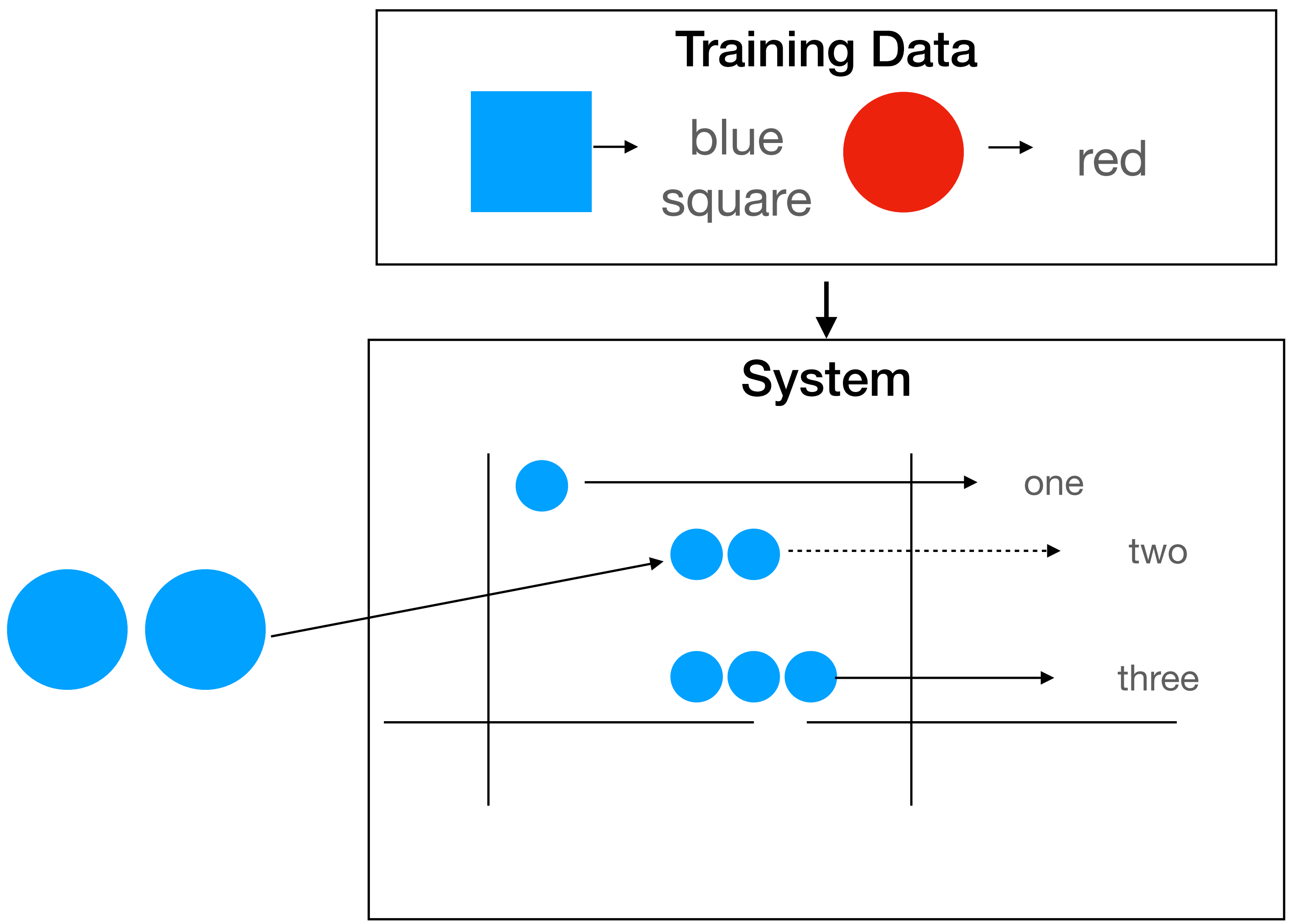
Not Sufficient: Models that don’t meet our definition can still succeed



Issue #2: “Unseen” is not well defined when we are working with distributed representations

Evaluating compositionality via behavior

Not Sufficient: Models that don’t meet our definition can still succeed



Issue #2: “Unseen” is not well defined when we are working with distributed representations

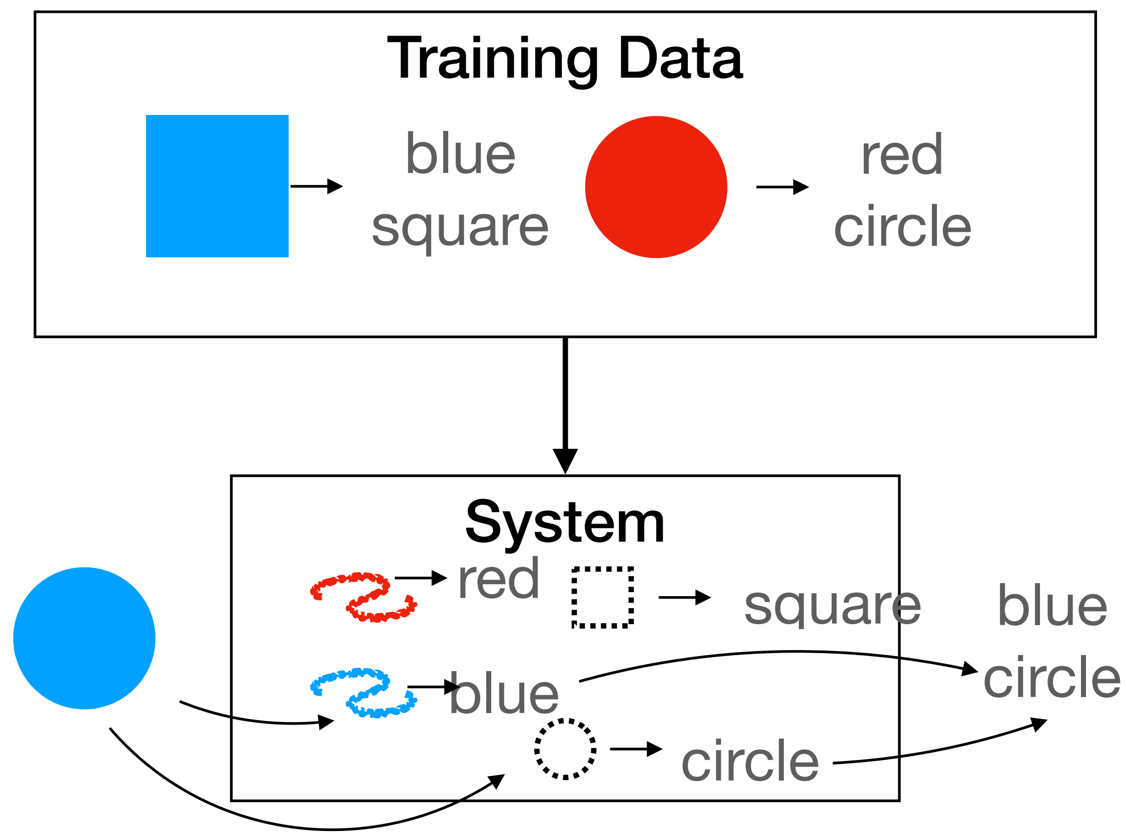
“In between” is not the same as “composed of”

Evaluating compositionality via behavior

Not Necessary: Models that meet our definition could still fail

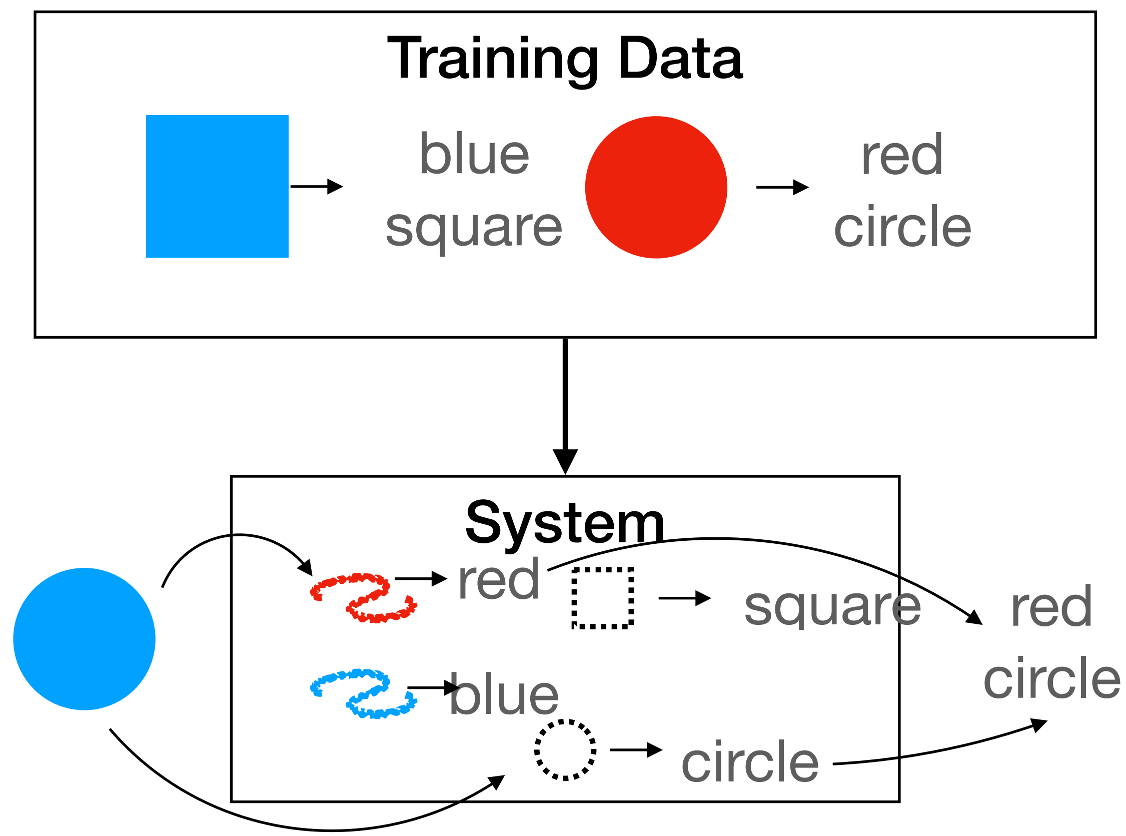
Evaluating compositionality via behavior

Not Necessary: Models that meet our definition could still fail



Evaluating compositionality via behavior

Not Necessary: Models that meet our definition could still fail

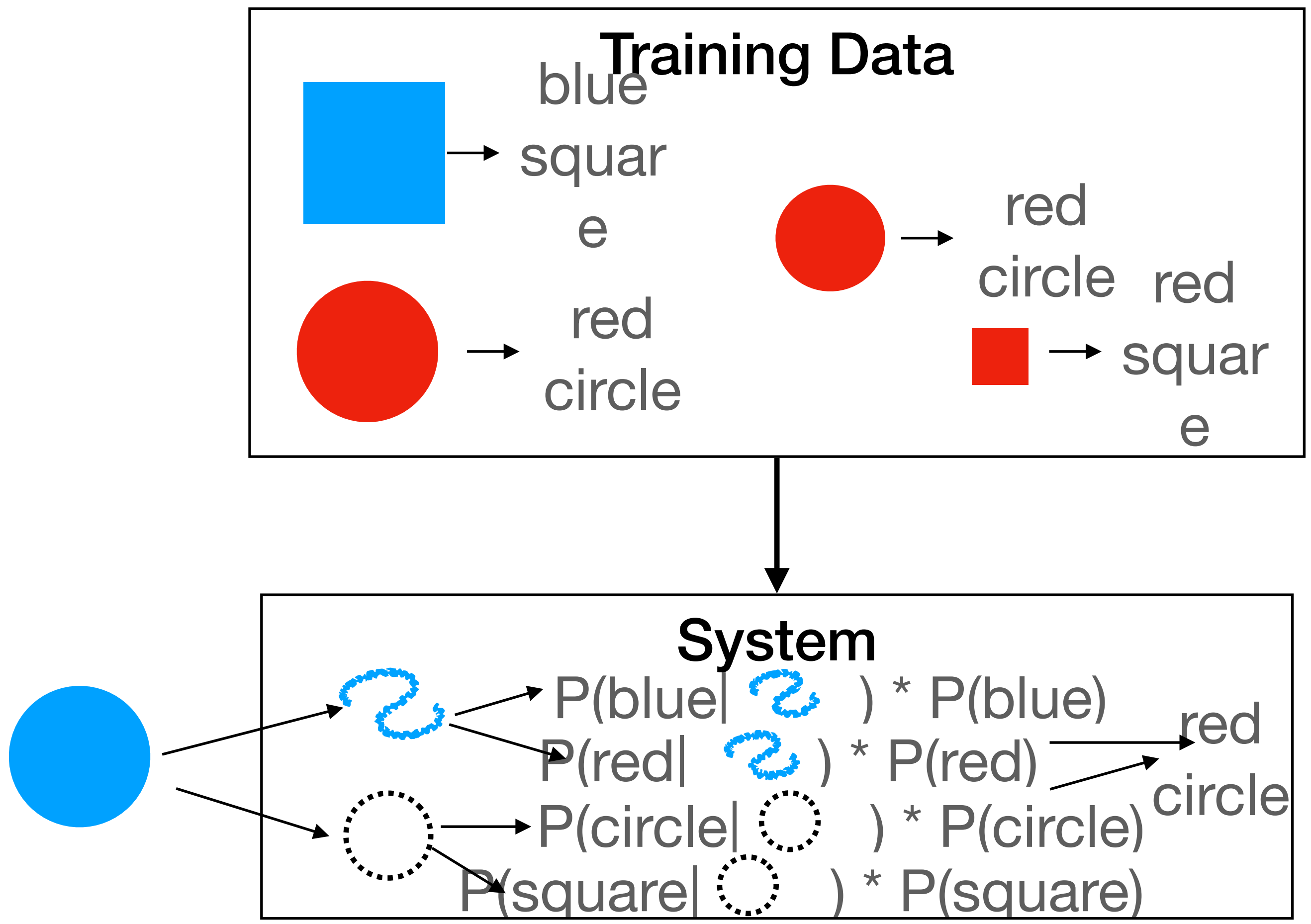


Issue #1: Compositional systems are allowed to make mistakes!

Bad visual perception does not entail "not compositional"

Evaluating compositionality via behavior

Not Necessary: Models that meet our definition could still fail



Issue #2: Compositional systems are allowed to be probabilistic!

Priors can (and often do) outweigh evidence, even in symbolic systems.

Structured Compositional Concepts

- Two questions:

    1. ~~Can~~ Do NNs *learn to implement* such a definition?

    2. If so, how would we know?

Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. [TACL 2022]

Jason Wei, Dan Garrette, Tal Linzen and Ellie Pavlick. Frequency Effects on Syntactic Rule Learning in Transformers. [EMNLP 2021]

Charles Lovering, **Rohan Jha**, Tal Linzen and Ellie Pavlick. Predicting Inductive Biases of Pretrained Models. [ICLR 2021]

Aaron Traylor, Roman Feiman and Ellie Pavlick. AND does not mean OR: Using Formal Languages to Study Language Models' Representations. [ACL 2021]

Roma Patel and Ellie Pavlick. Mapping Language Models to Grounded Conceptual Spaces. [ICLR 2022]

Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. [TACL 2022]

Jason Wei, Dan Garrette, Tal Linzen and Ellie Pavlick. Frequency Effects on Syntactic Rule Learning in Transformers. [EMNLP 2021]

Charles Lovering, **Rohan Jha**, Tal Linzen and Ellie Pavlick. Predicting Inductive Biases of Pretrained Models. [ICLR 2021]

Aaron Traylor, Roman Feiman and Ellie Pavlick. AND does not mean OR: Using Formal Languages to Study Language Models' Representations. [ACL 2021]

Roma Patel and Ellie Pavlick. Mapping Language Models to Grounded Conceptual Spaces. [ICLR 2022]

Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. [TACL 2022]

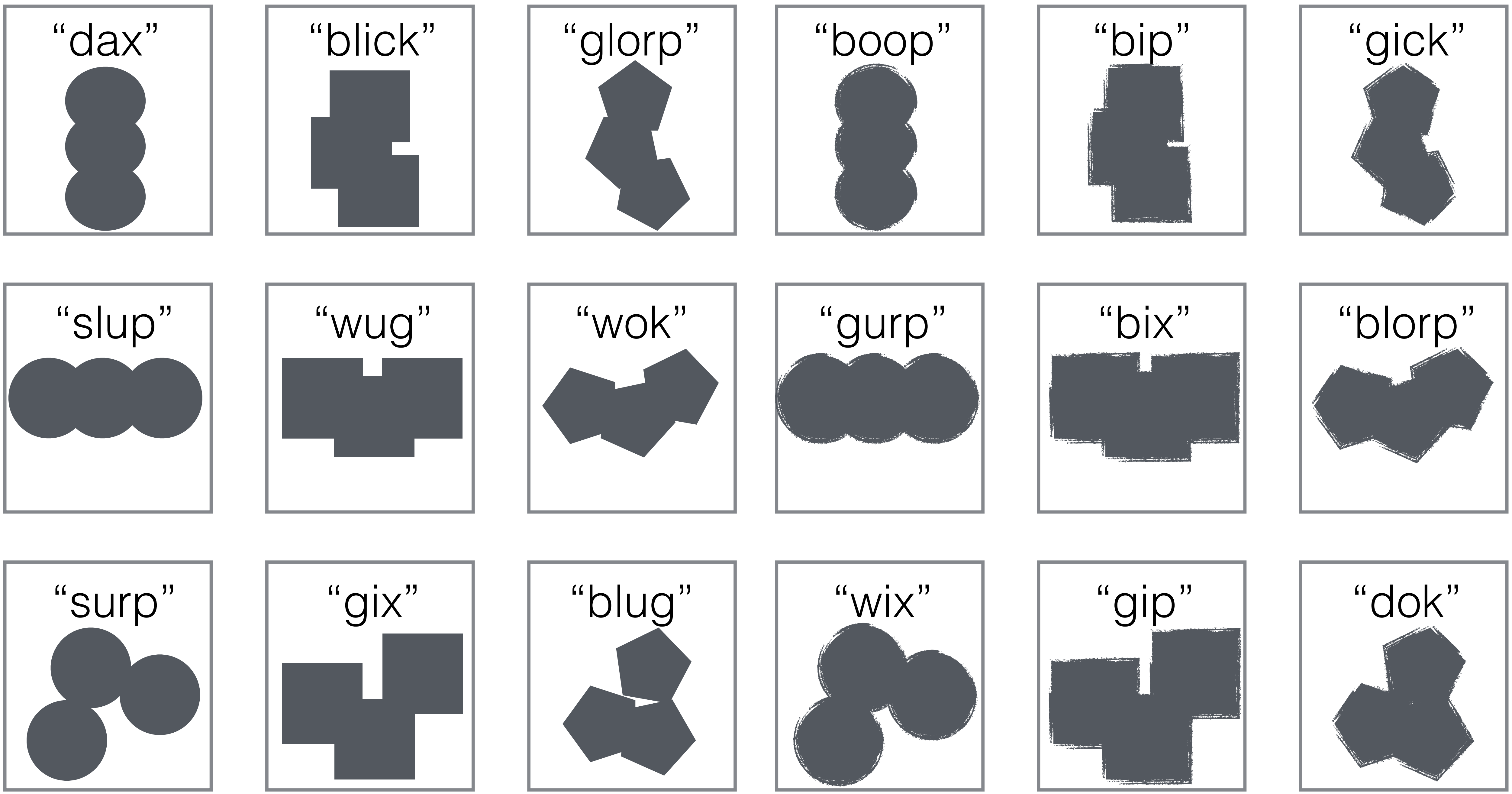
Jason Wei, Dan Garrette, Tal Linzen and Ellie Pavlick. Frequency Effects on Syntactic Rule Learning in Transformers. [EMNLP 2021]

Charles Lovering, **Rohan Jha**, Tal Linzen and Ellie Pavlick. Predicting Inductive Biases of Pretrained Models. [ICLR 2021]

Aaron Traylor, Roman Feiman and Ellie Pavlick. AND does not mean OR: Using Formal Languages to Study Language Models' Representations. [ACL 2021]

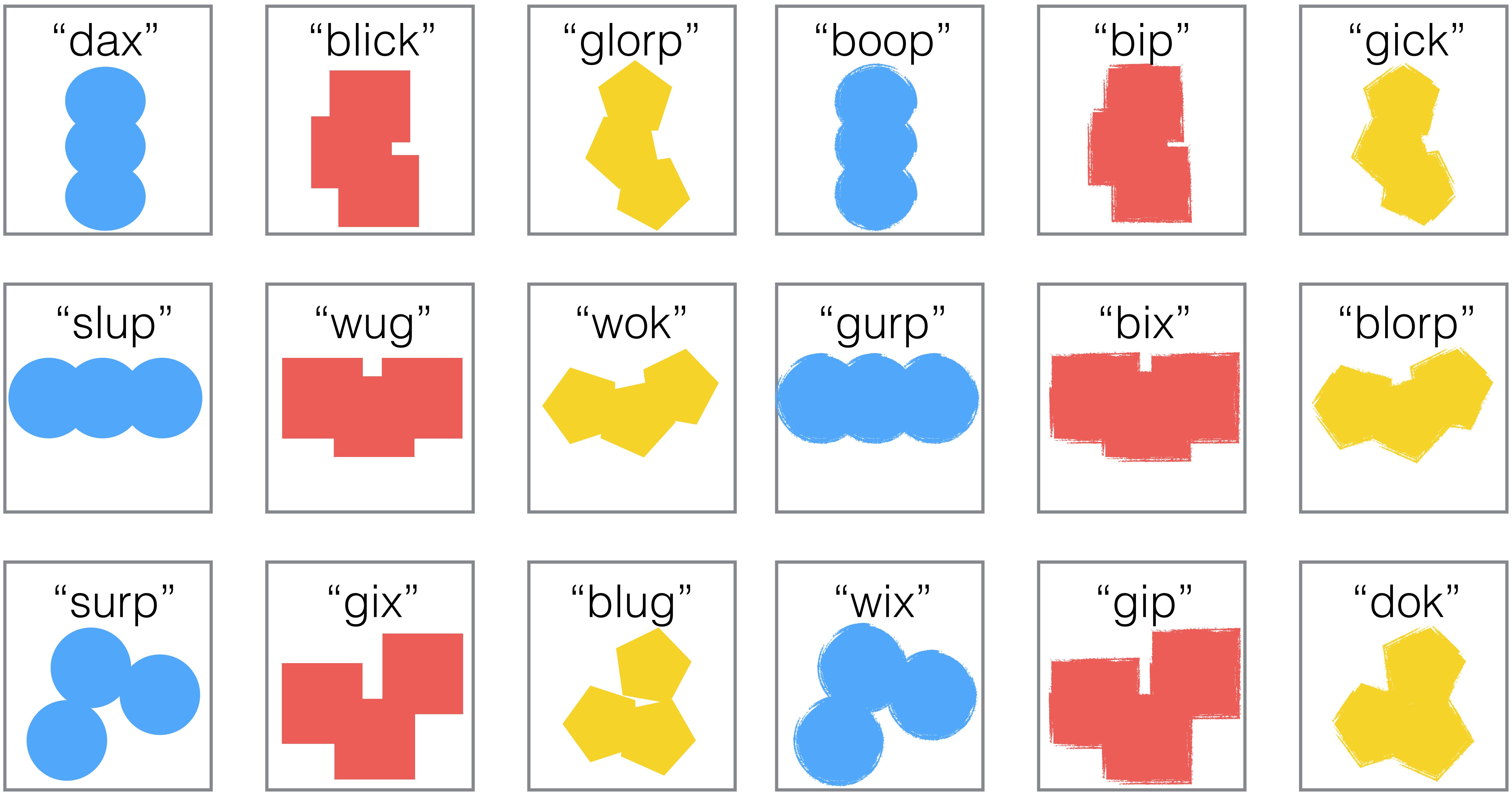
Roma Patel and Ellie Pavlick. Mapping Language Models to Grounded Conceptual Spaces. [ICLR 2022]

Task Setup: Labeling Simple Visual Concepts



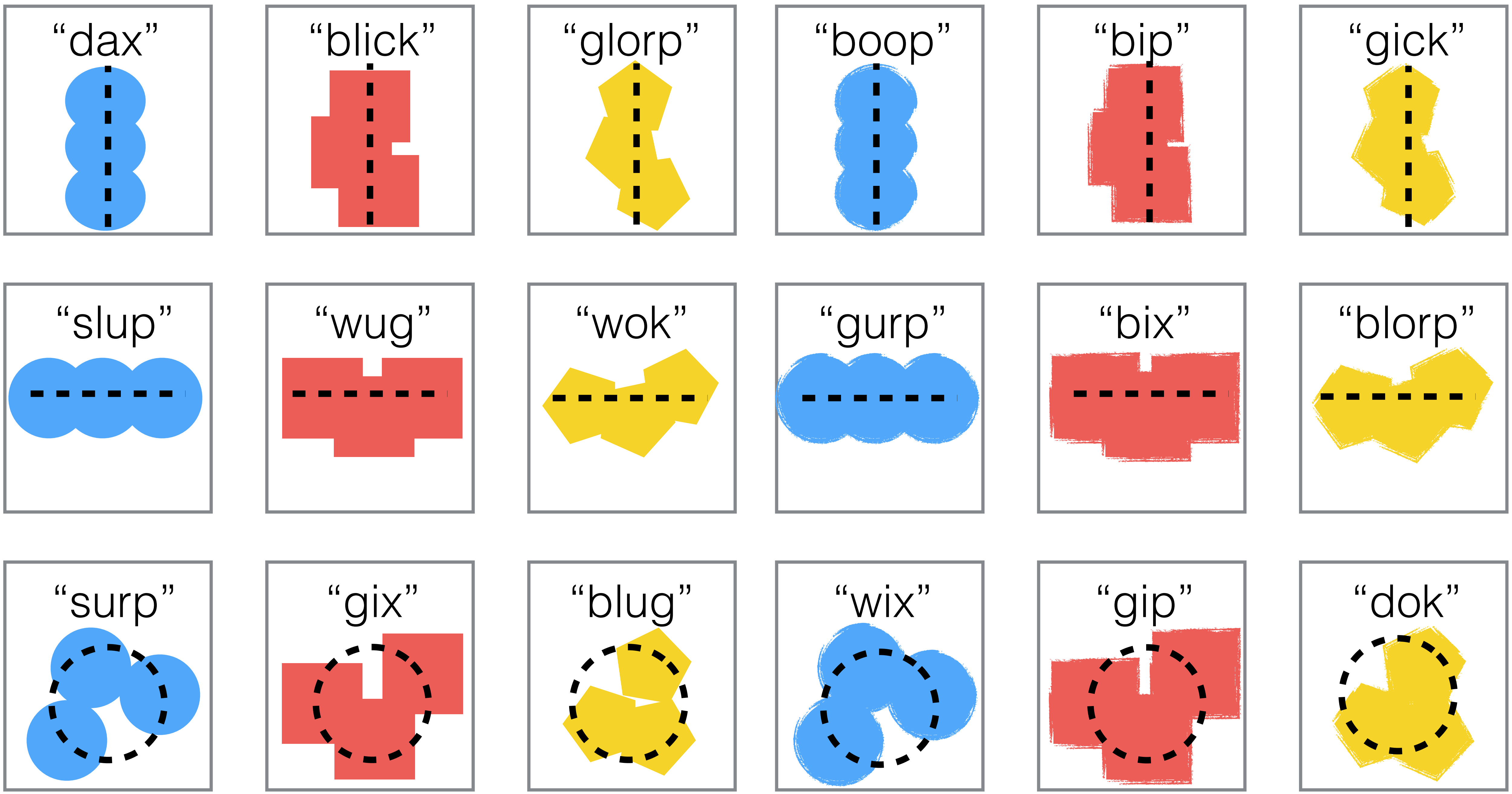
18 high level concepts composed from **8 basic concepts**

Task Setup: Labeling Simple Visual Concepts



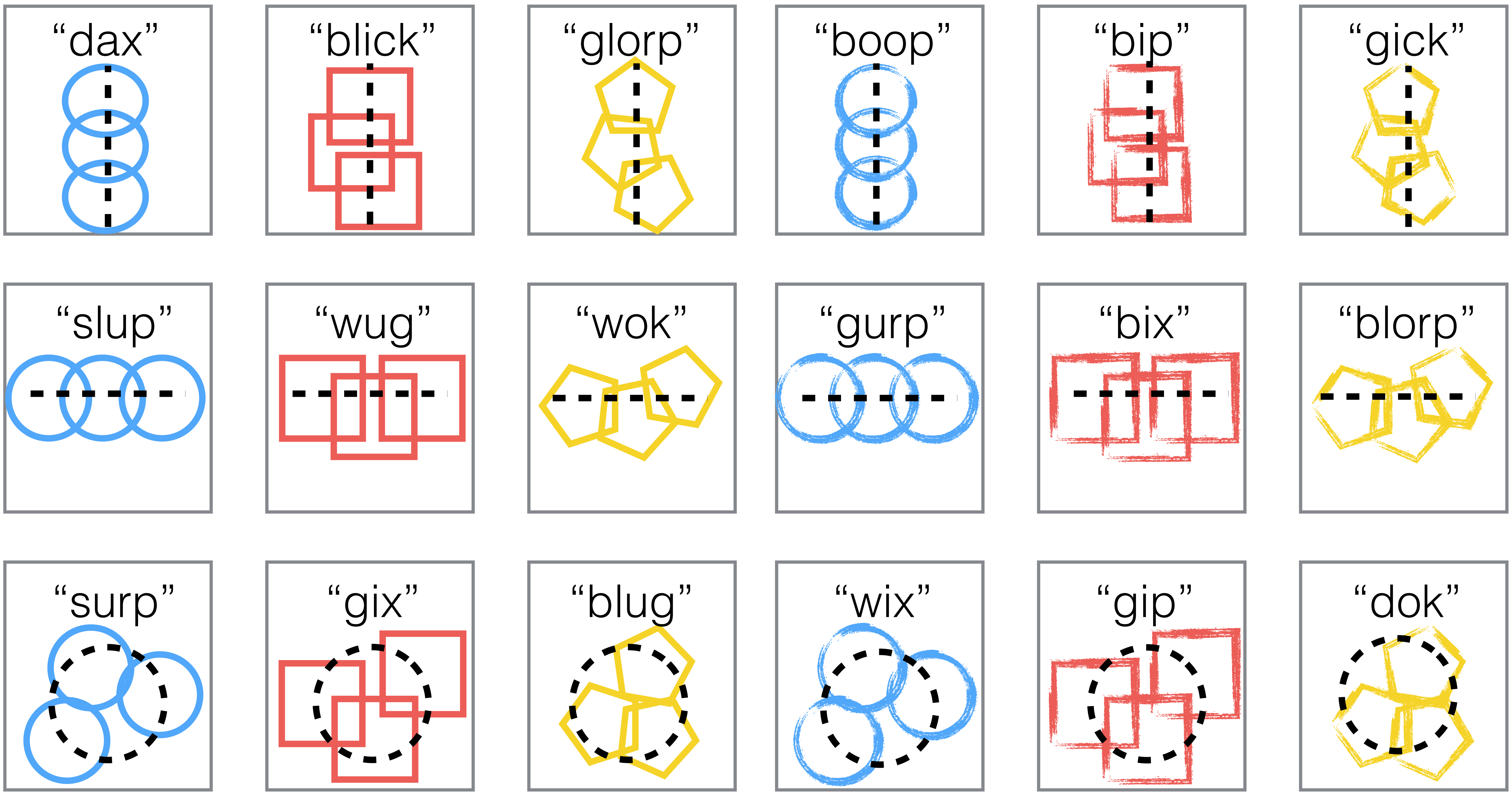
18 high level concepts = {shape}

Task Setup: Labeling Simple Visual Concepts



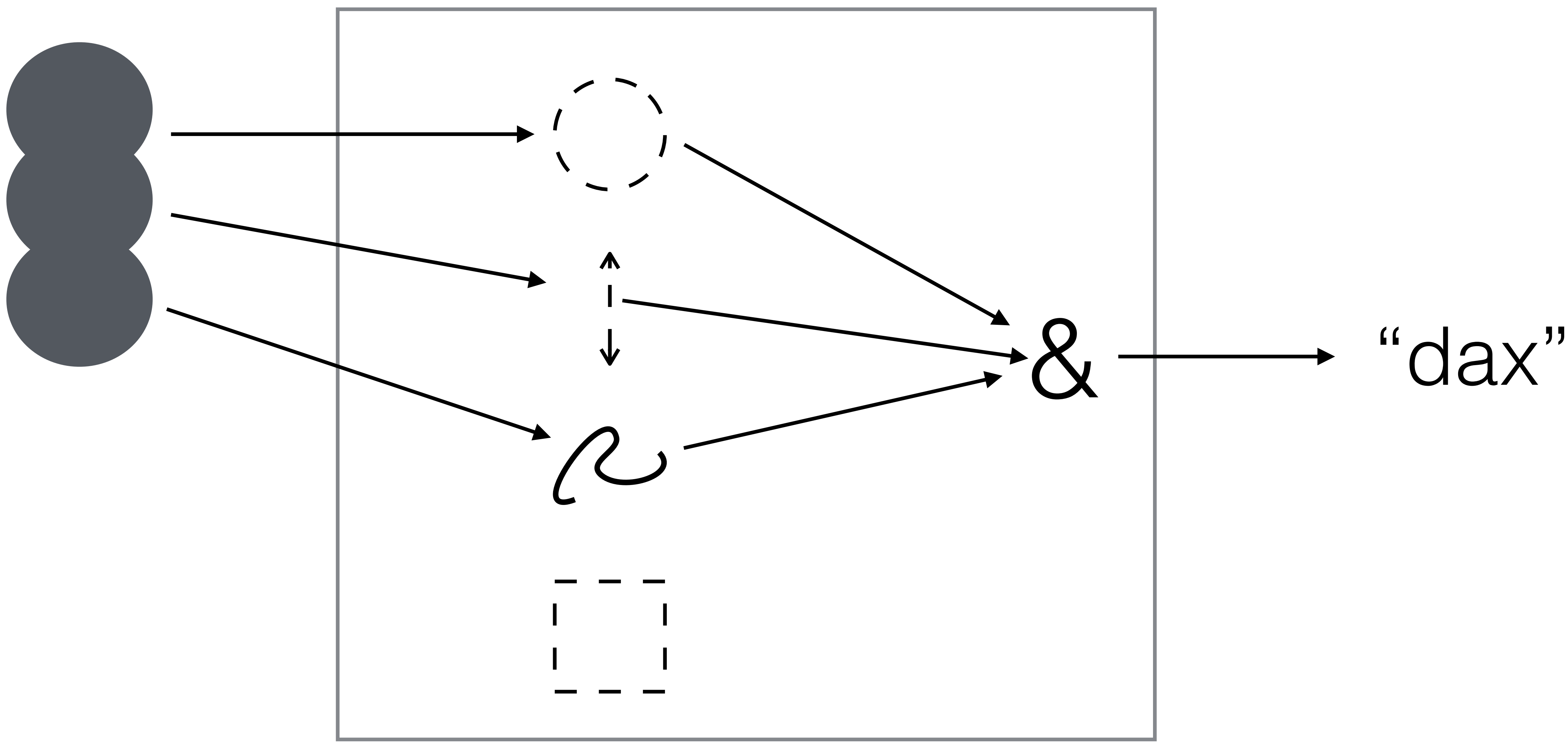
18 high level concepts = {shape} x {layout}

Task Setup: Labeling Simple Visual Concepts

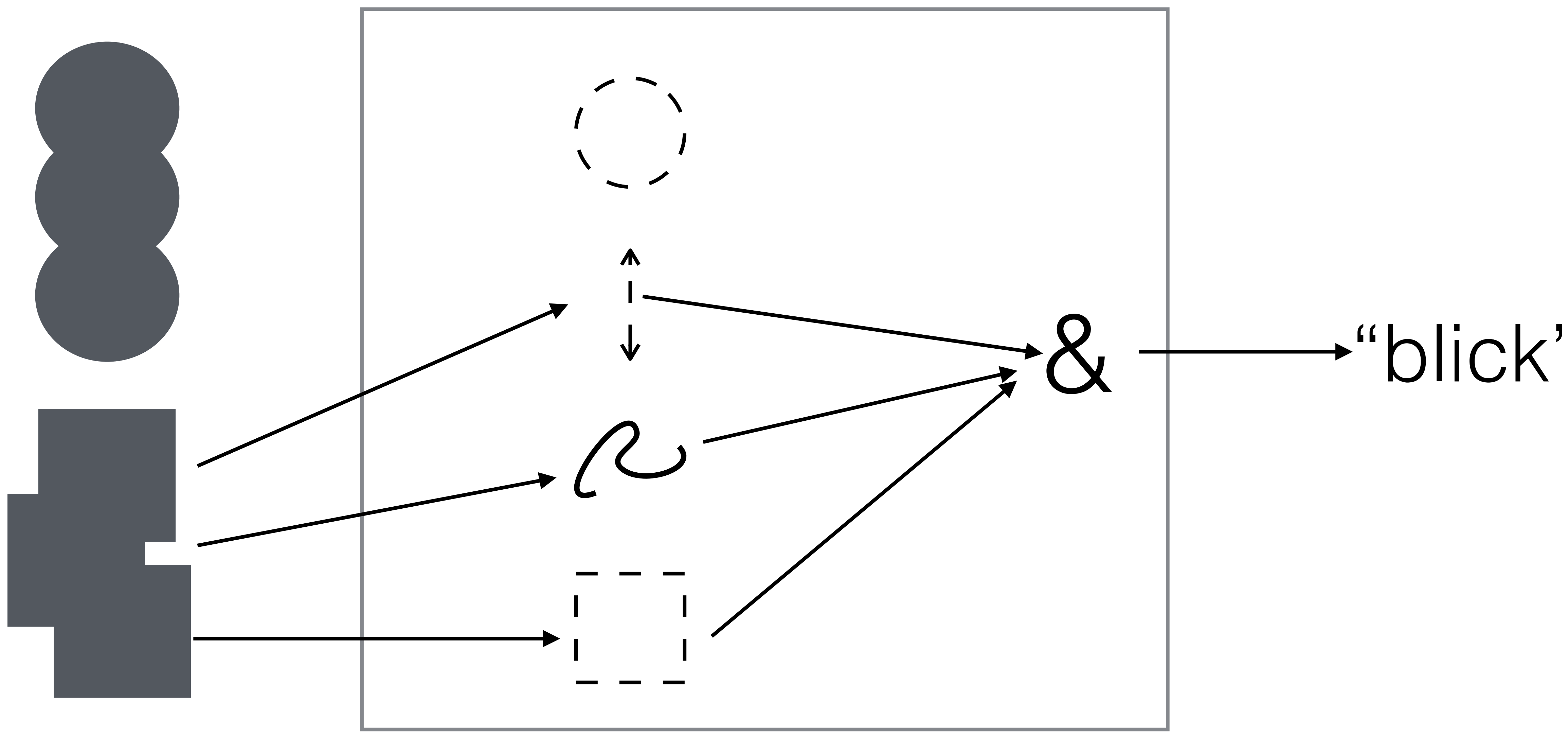


18 high level concepts = {shape} x {layout} x {stroke}

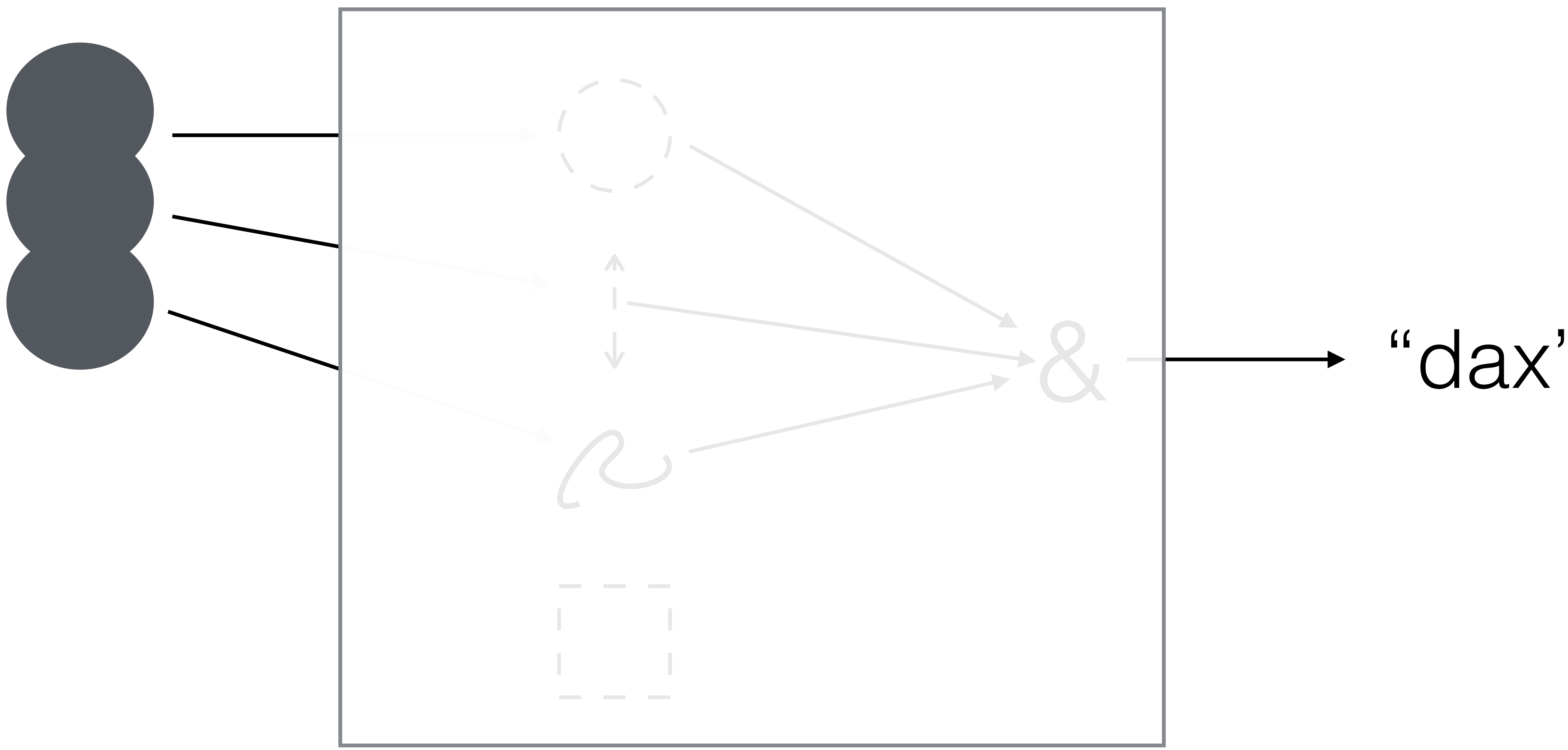
High-Level API



High-Level API

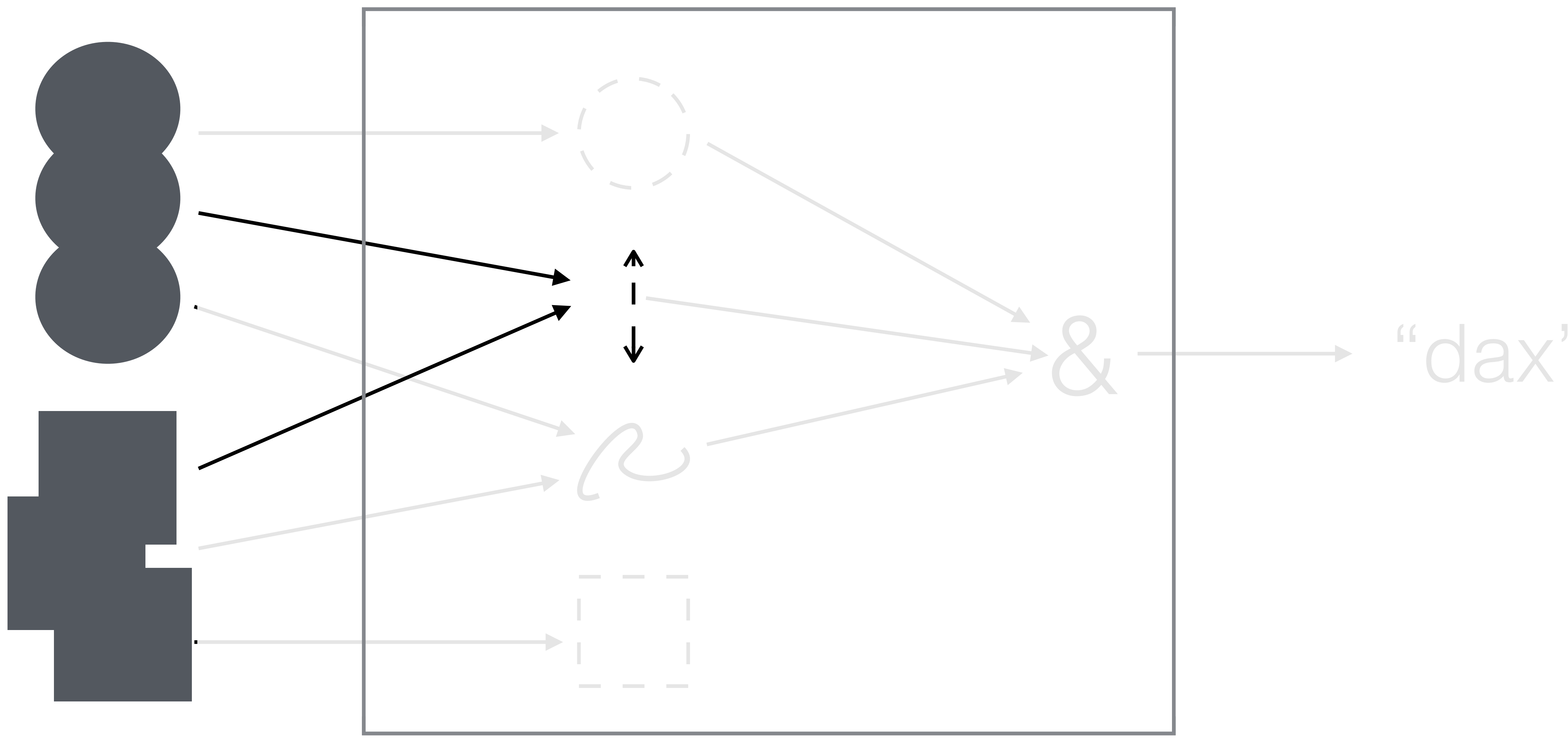


Requirement #1: Predictions are grounded



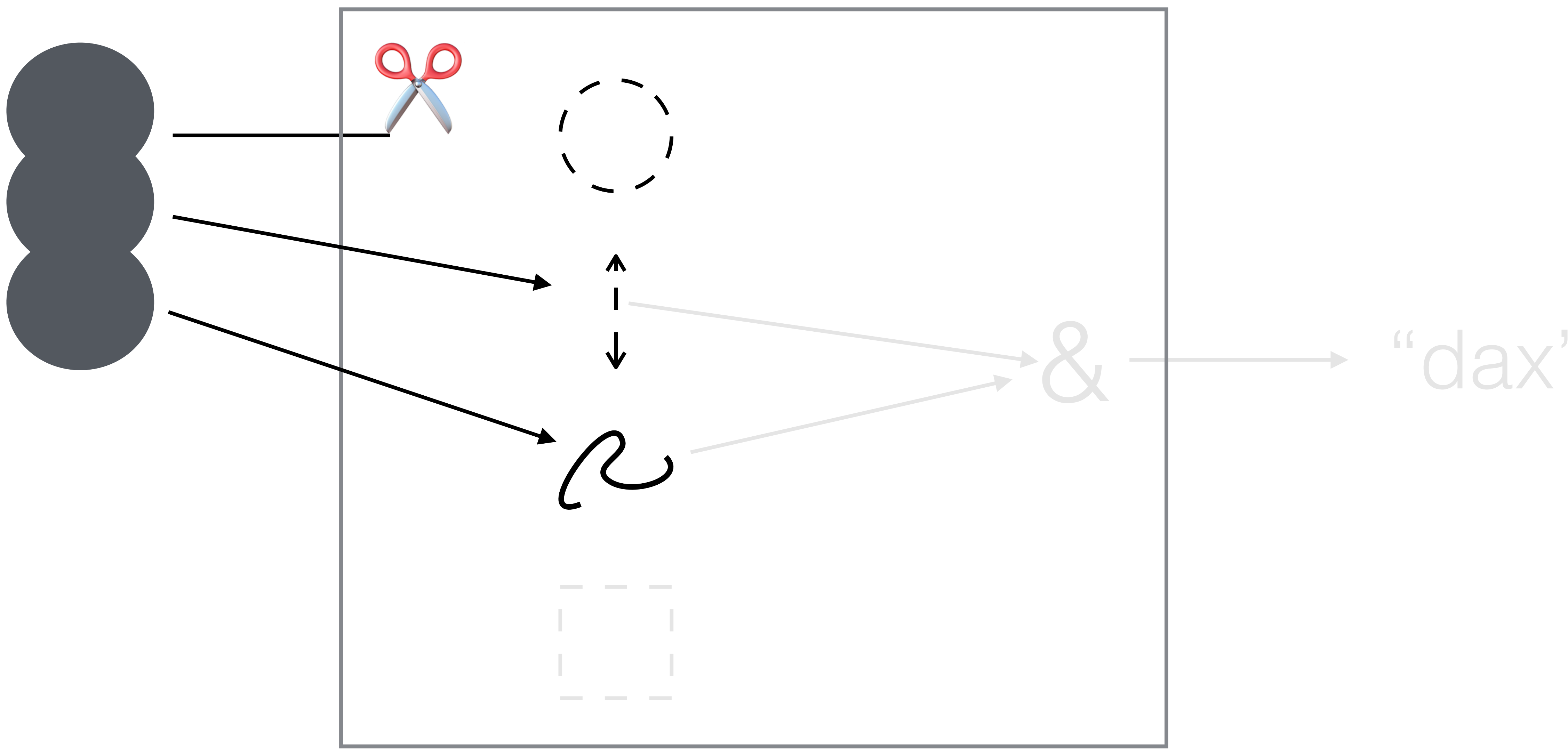
B: Concepts apply to things in the world
E: Concepts are public

Requirement #2: Concepts represent types



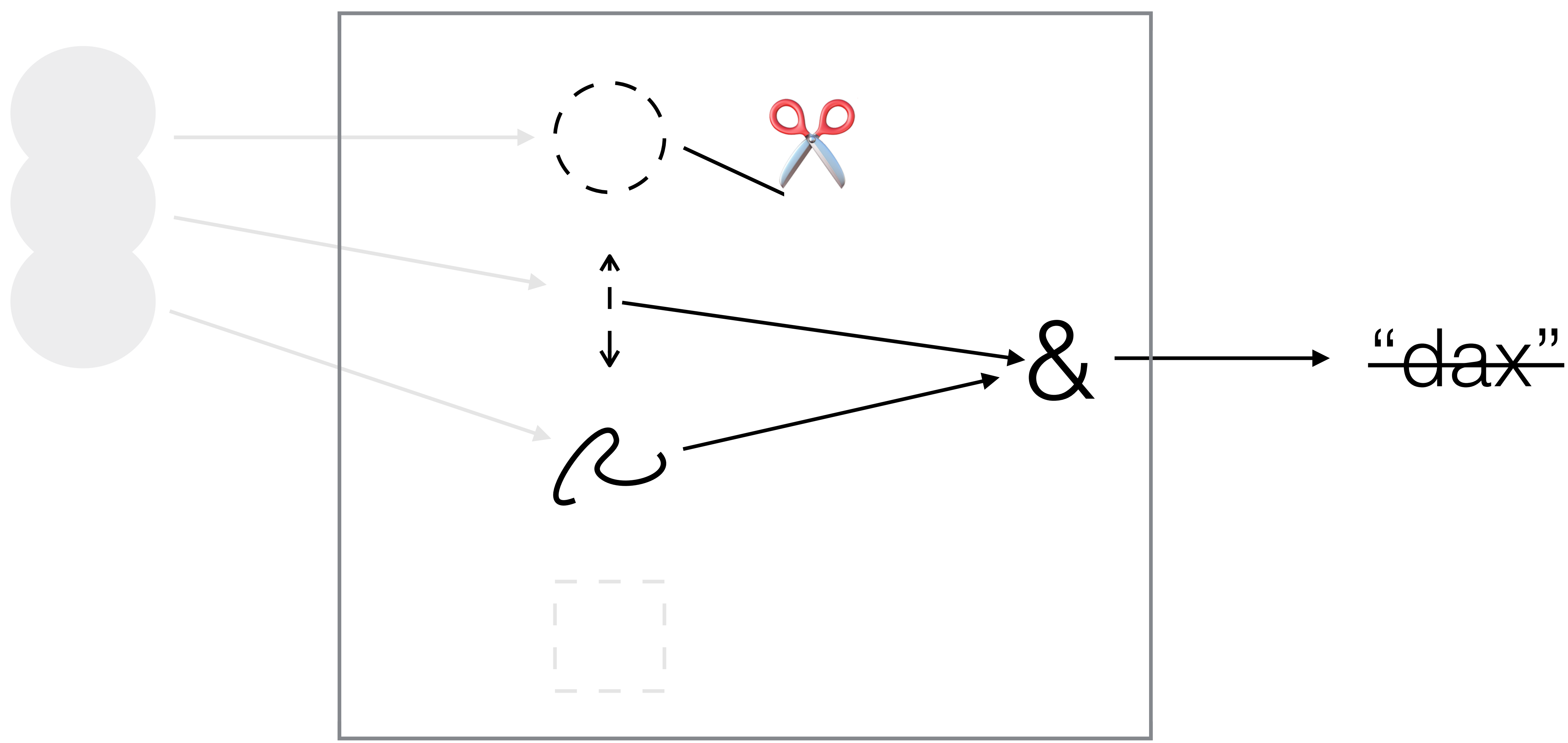
C: Constituency Structure: different tokens but a single type (Fodor&Phylyshn 1988)

Requirement #3: Concepts are modular



C: Constituency Structure: constituents obey rules of syntax; changes within a constituent should not have side effects. (Fodor&Phylyshn 1988)

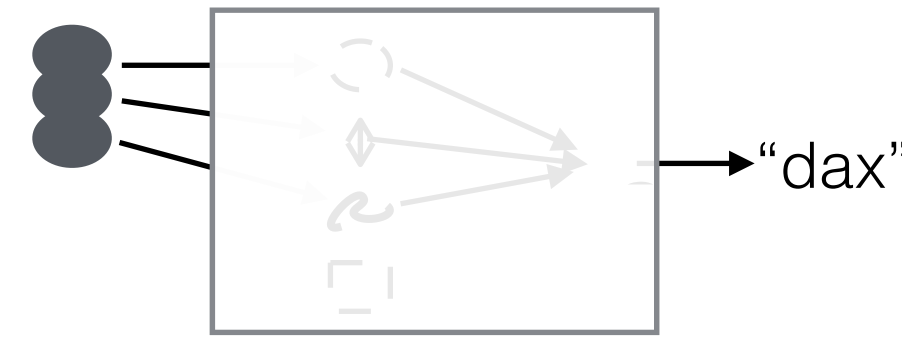
Requirement #4: Concepts are causal



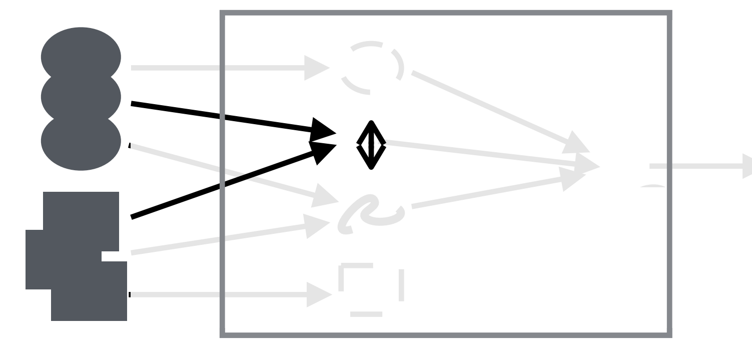
A: Function as Mental Causes and Effects

High-Level API

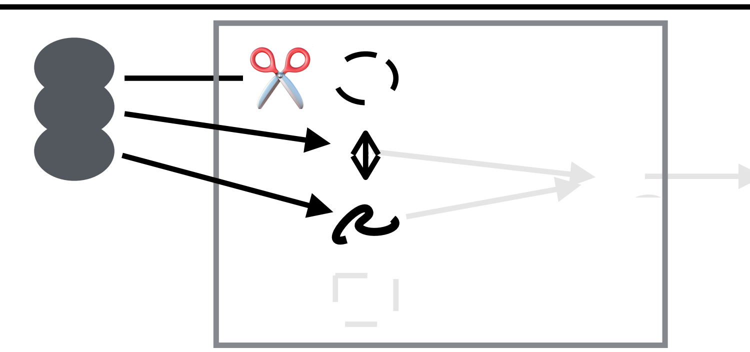
Predictions are
grounded



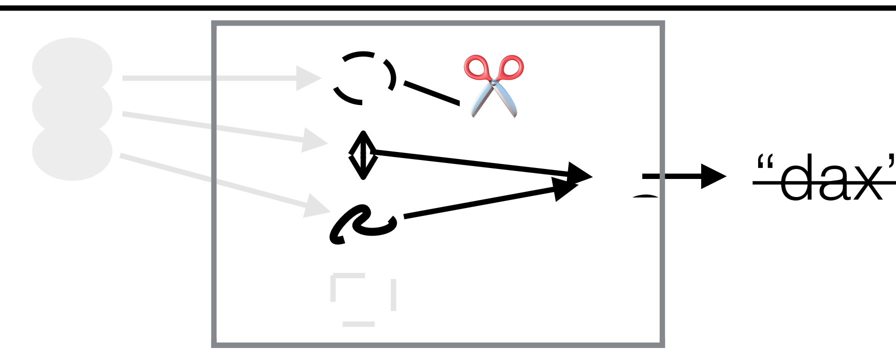
Concepts represent
types



Concepts are
modular



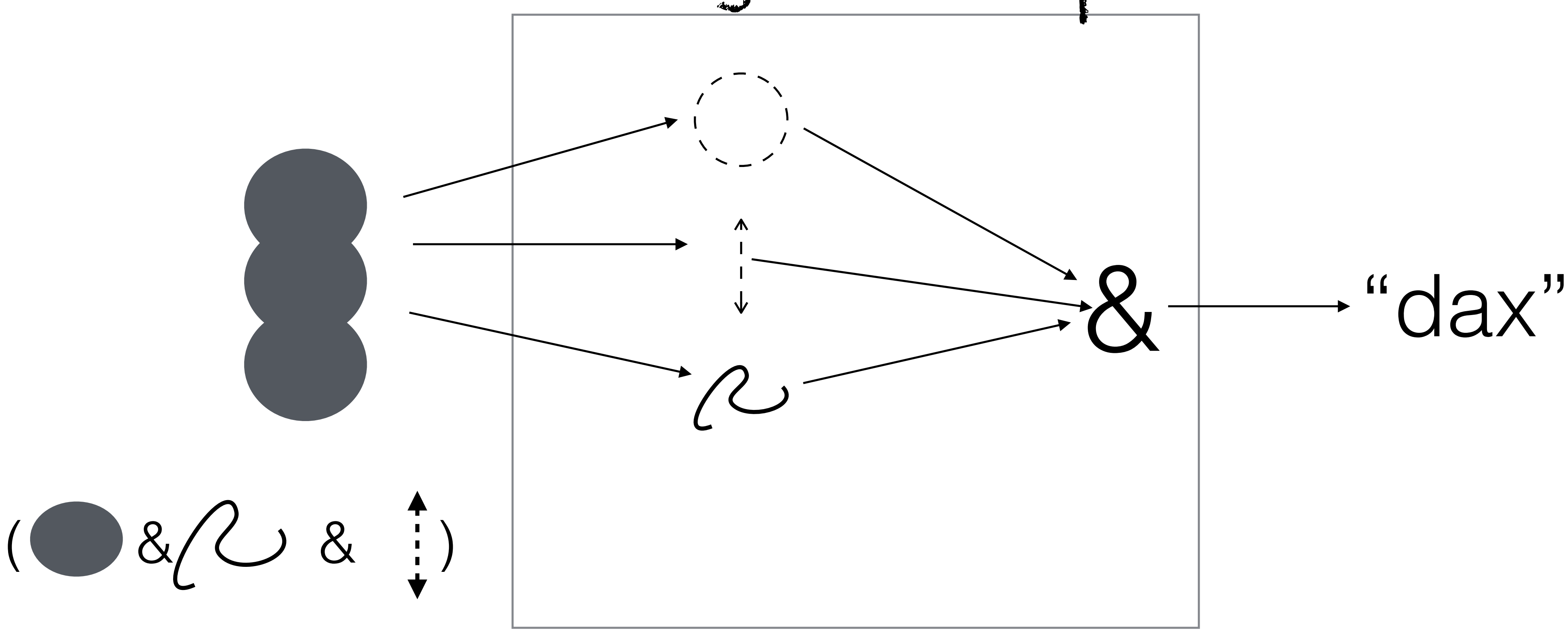
Concepts are
causal



Requirement #1:
Predictions are grounded

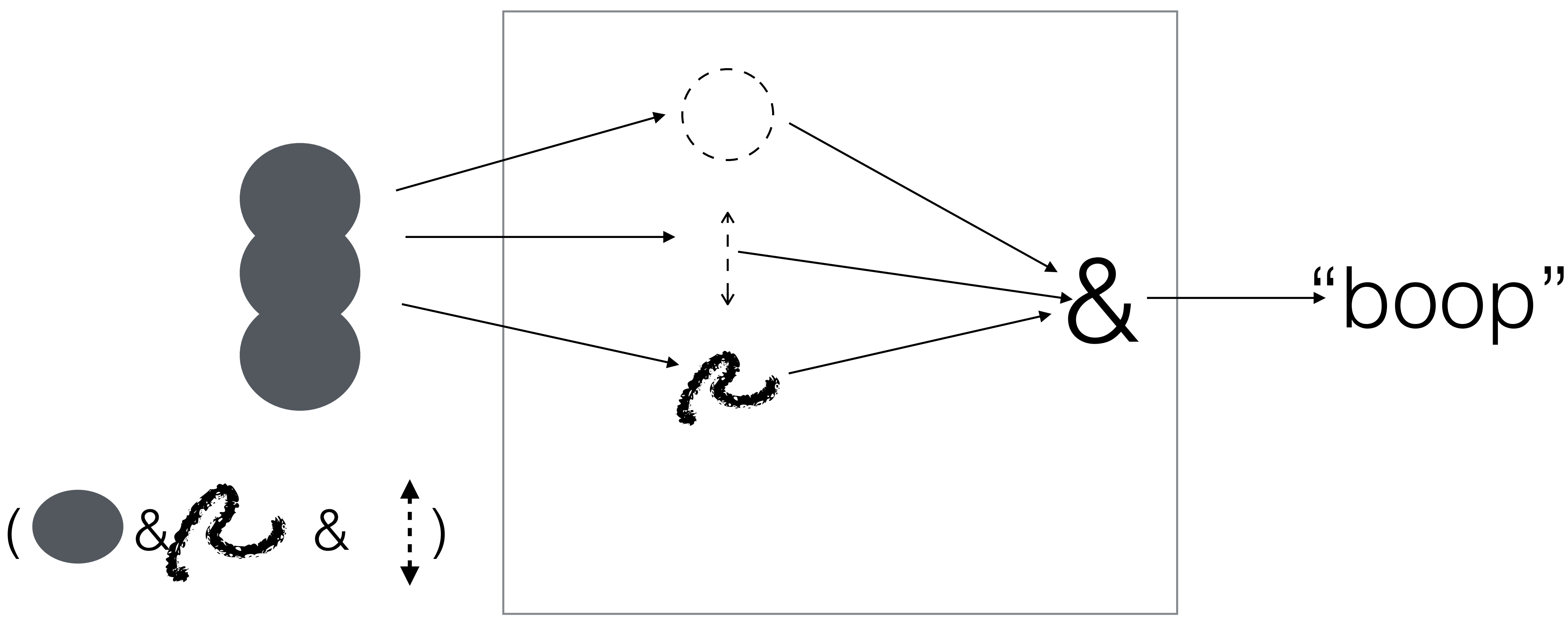
Requirement #1: Predictions are grounded

Changes in input lead to expected changes in output.

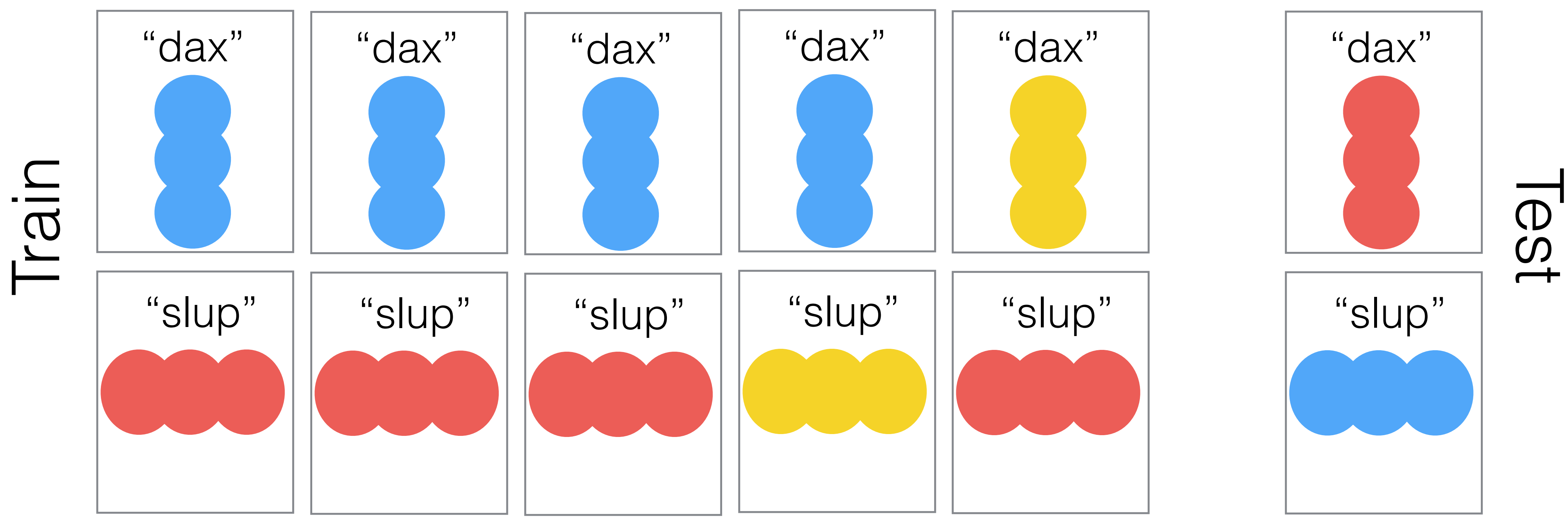


Requirement #1: Predictions are grounded

Evaluate using counterfactual minimal pairs

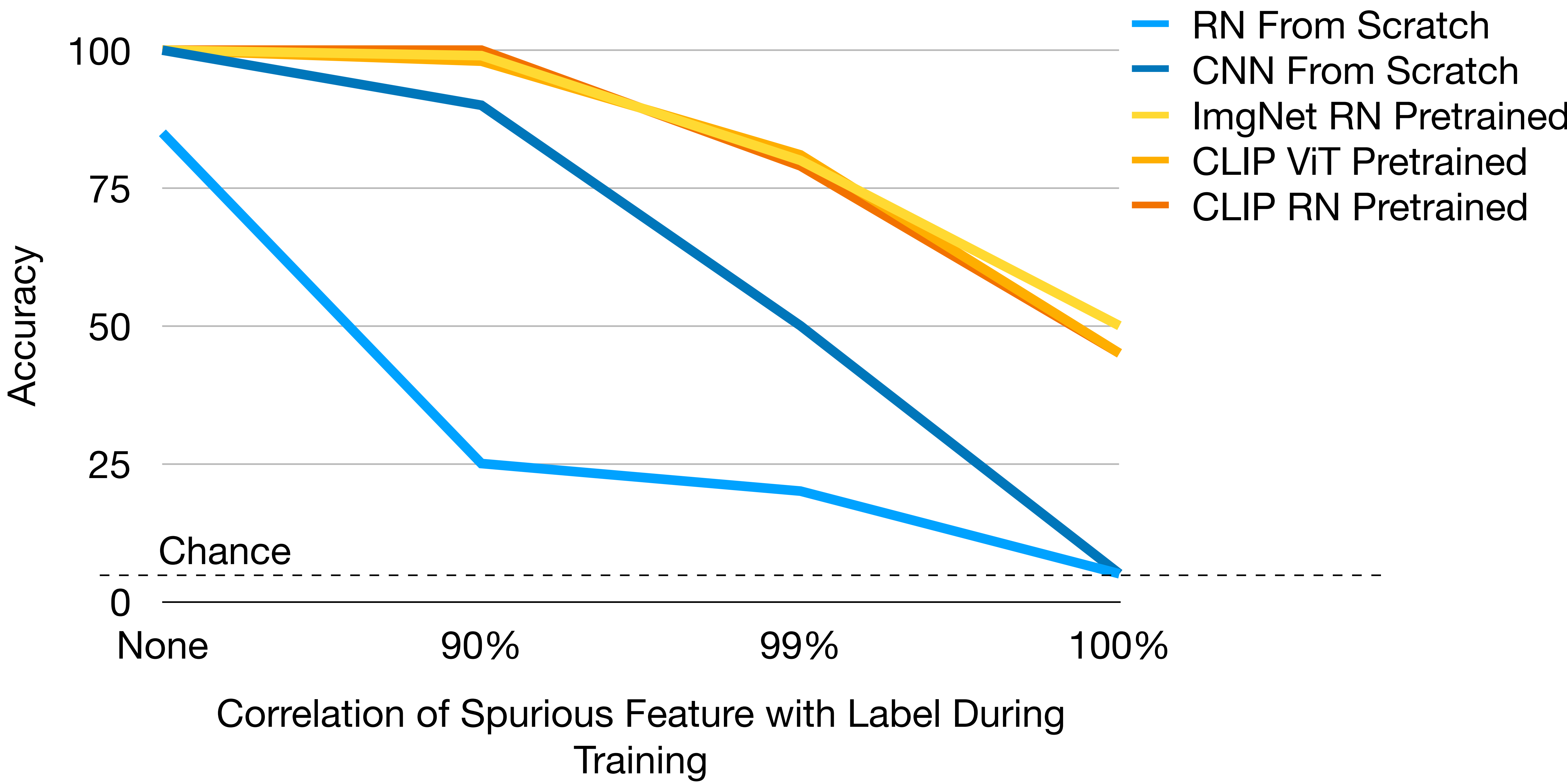


Requirement #1: Predictions are grounded

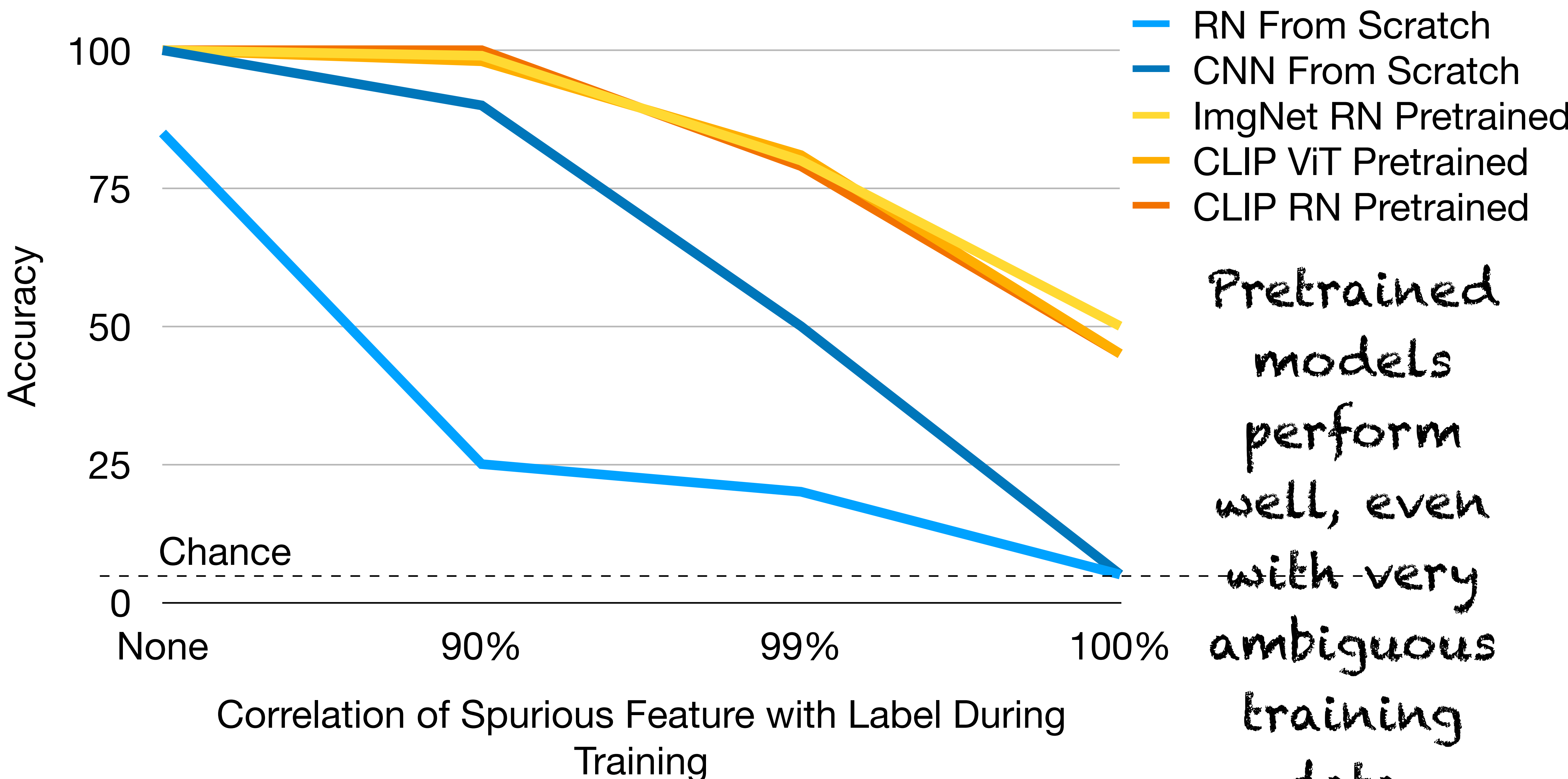


Introduce color as a correlated ("spurious") feature

Requirement #1: Predictions are grounded

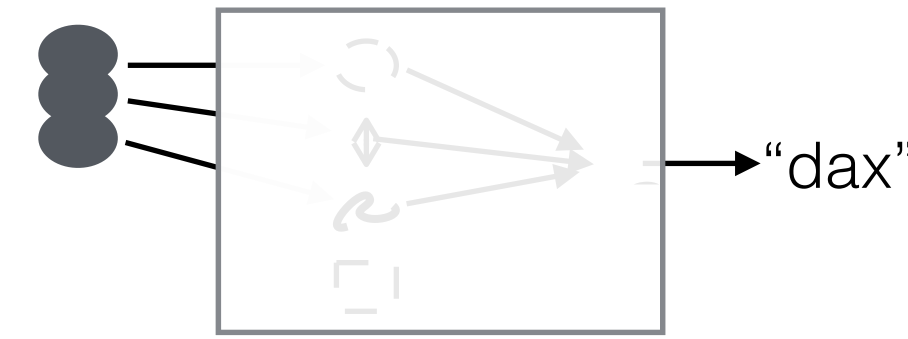
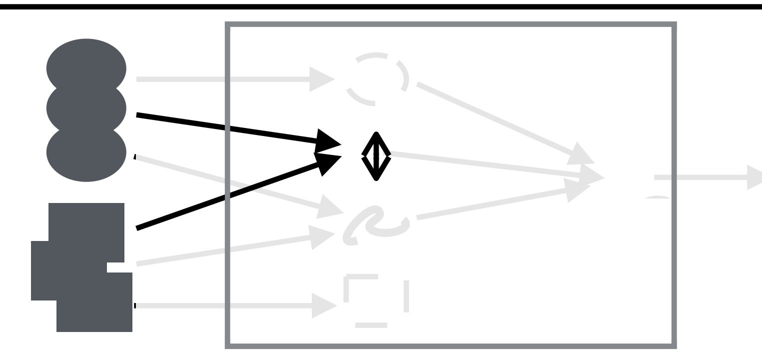
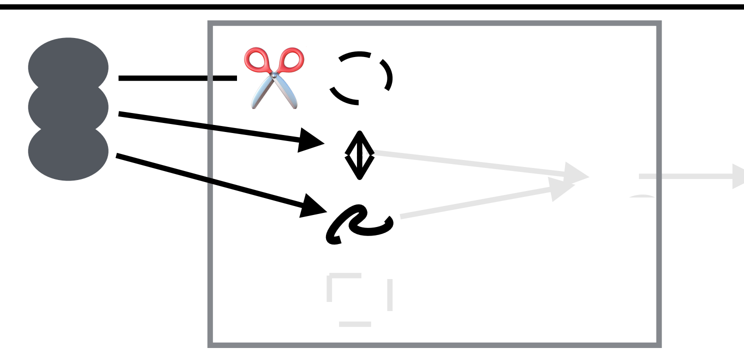
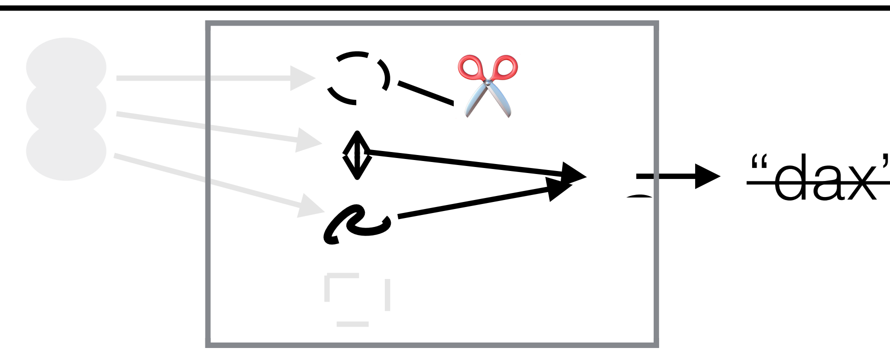


Requirement #1: Predictions are grounded



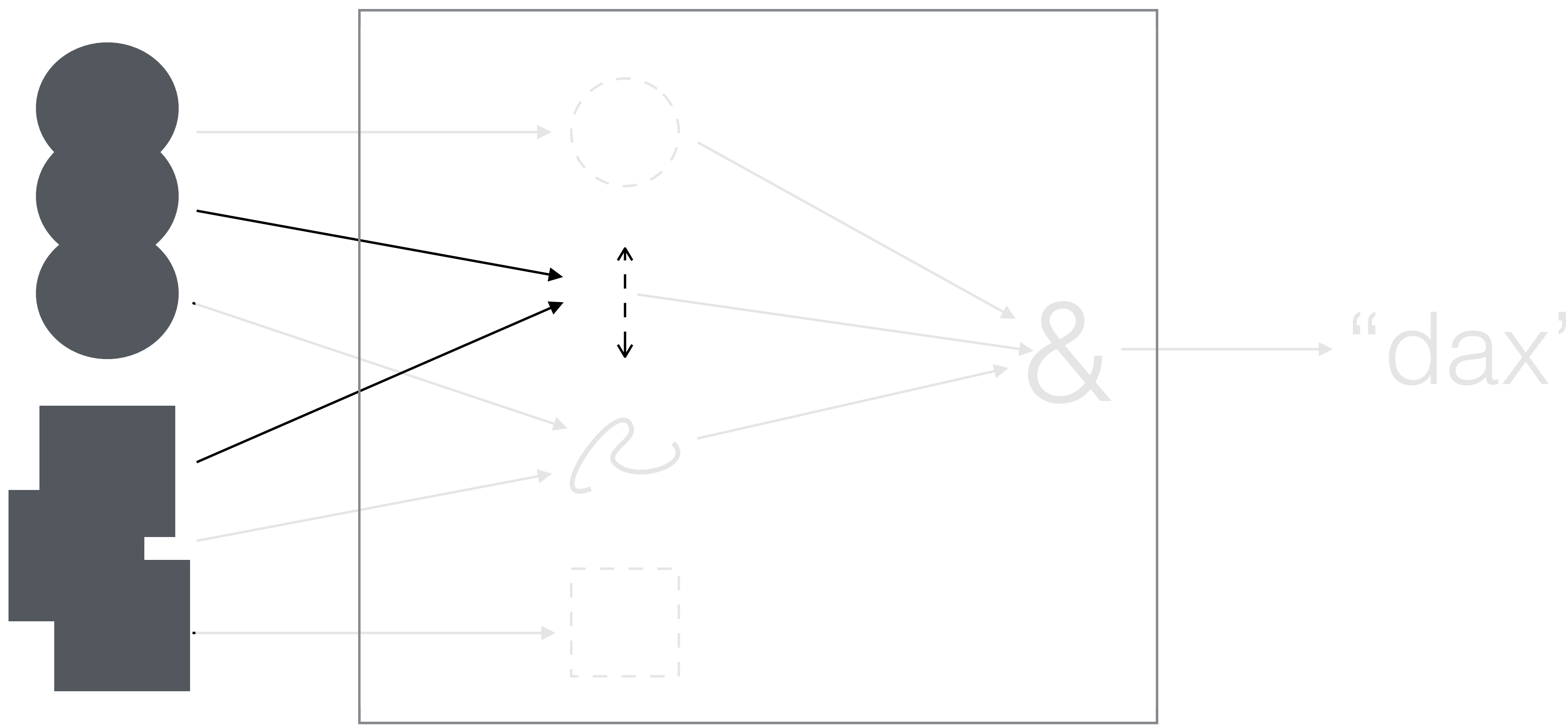
Pretrained models perform well, even with very ambiguous training data.

High-Level API

✅	Predictions are **grounded**	
	Concepts represent **types**	
	Concepts are **modular**	
	Concepts are **causal**	

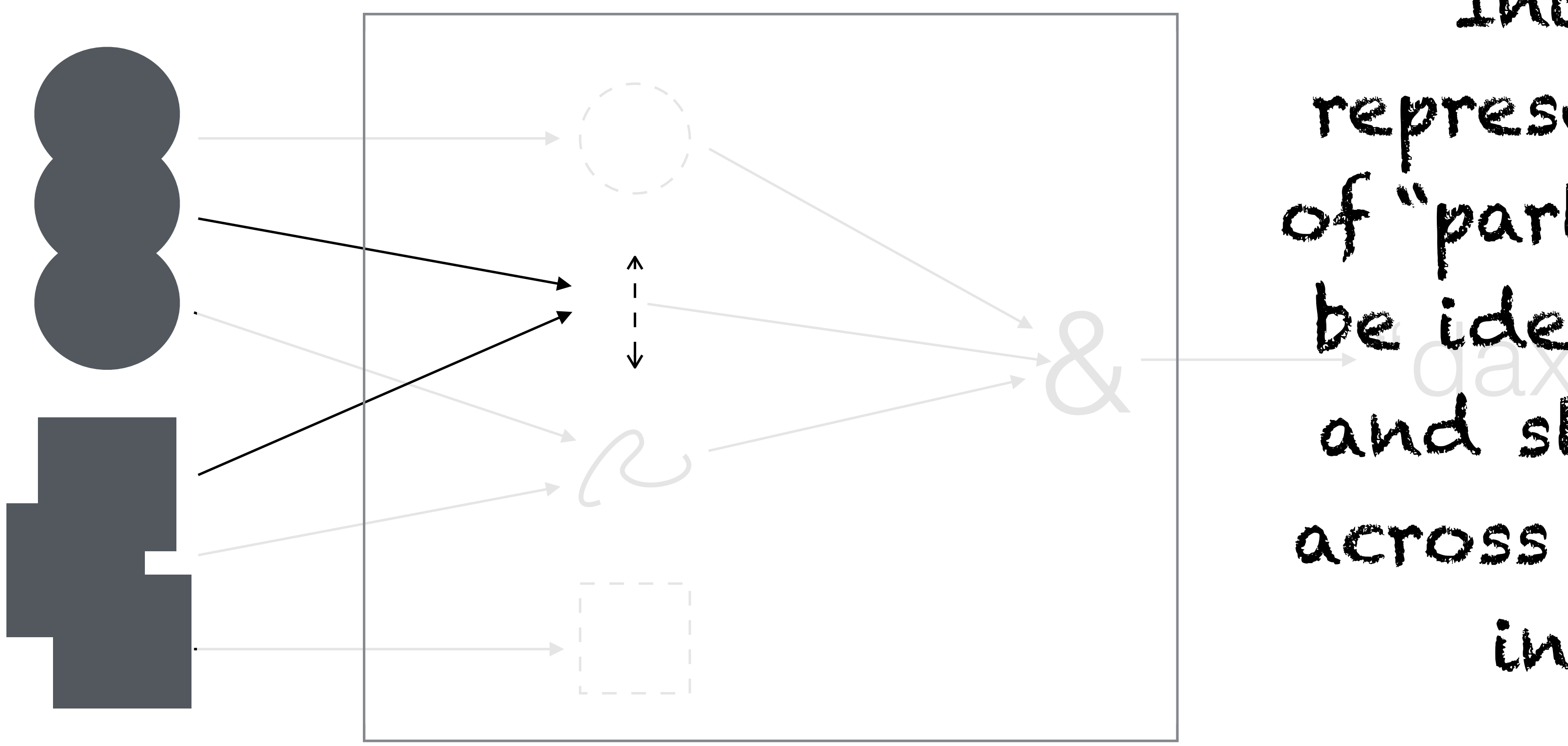
Requirement #2:
Concepts represent types

Requirement #2: Concepts represent types



“The ability to produce/understand some sentences is *intrinsically* connected to the ability to produce/understand certain others… [they] *must be made of the same parts*.”
(Fodor&Pylyshyn, 1988)

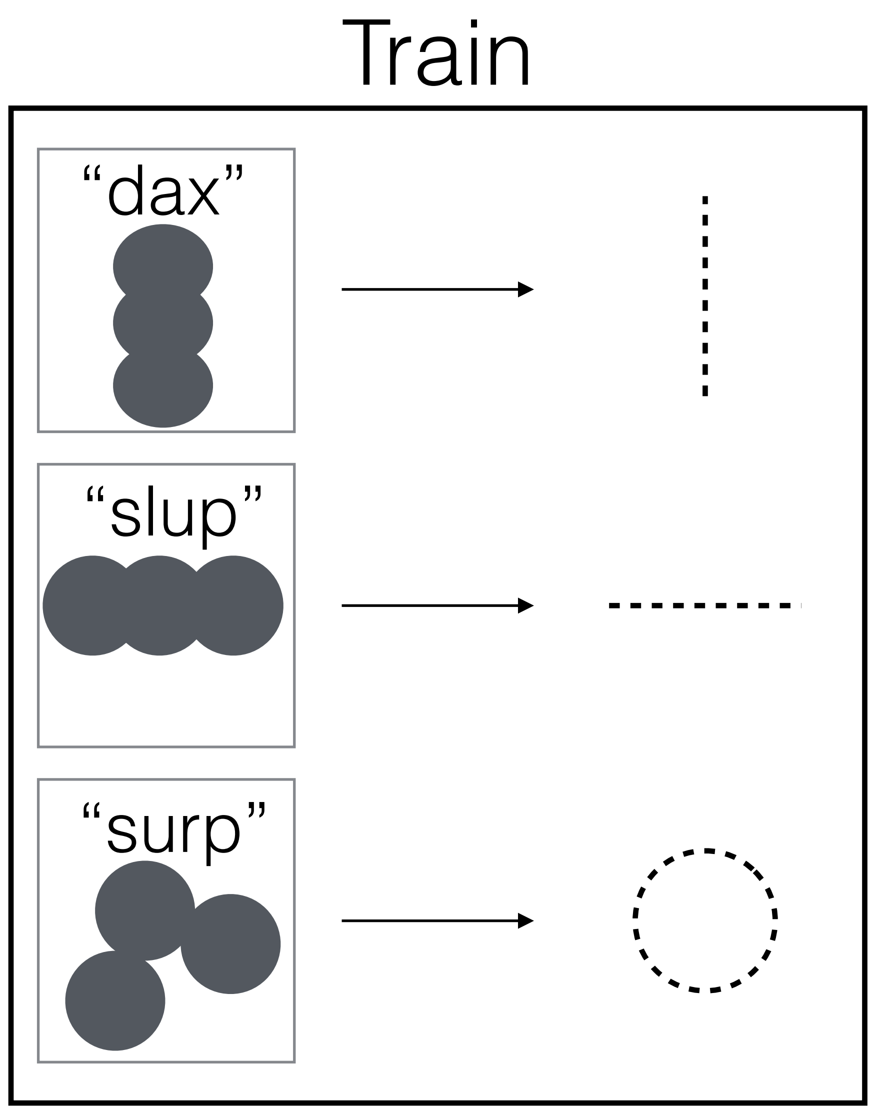
Requirement #2: Concepts represent types



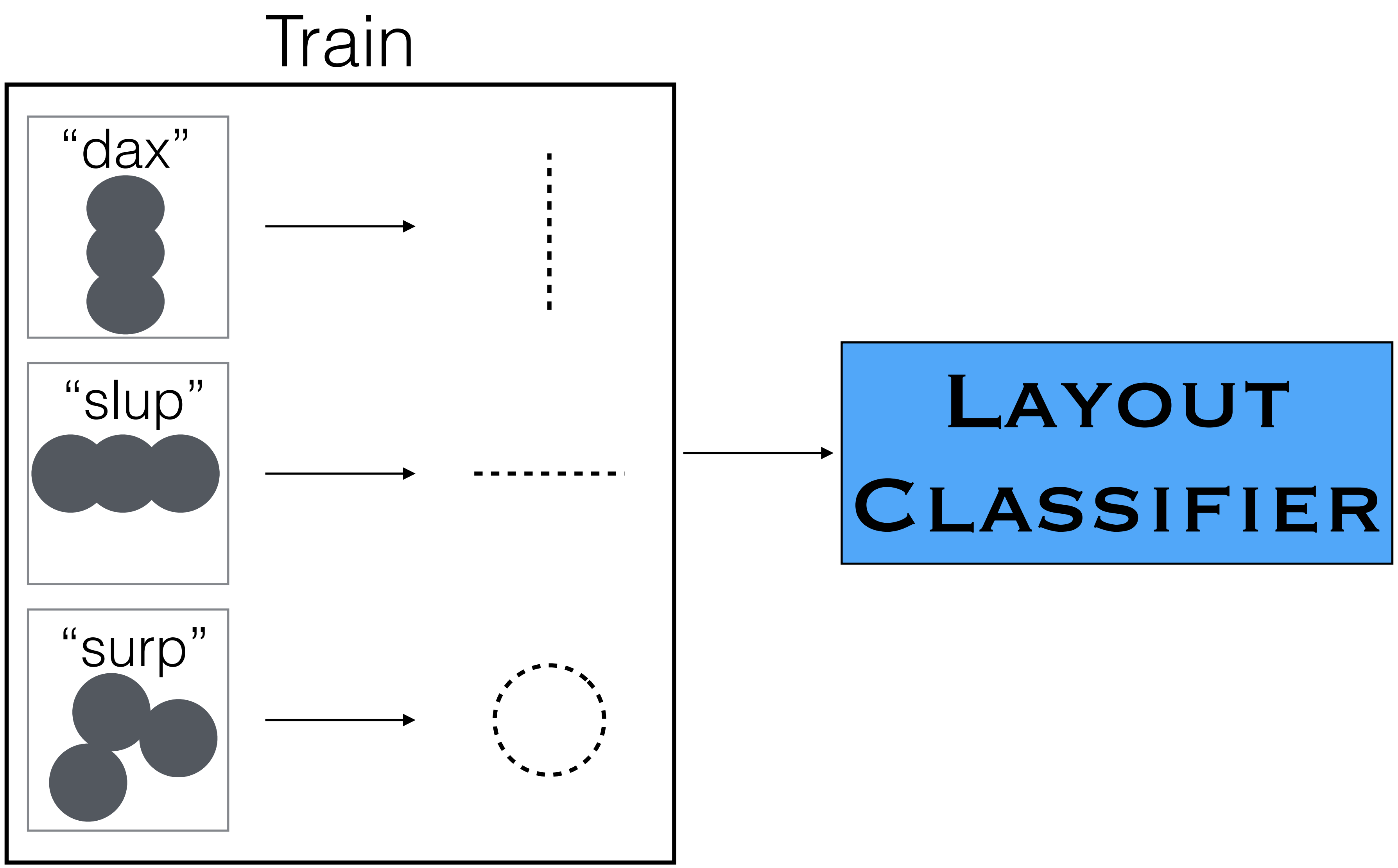
Internal representations of "parts" should be identifiable, and stable(ish) across different inputs.

"The ability to produce/understand some sentences is *intrinsically* connected to the ability to produce/understand certain others… [they] *must be made of the same parts*."
(Fodor&Pylyshyn, 1988)

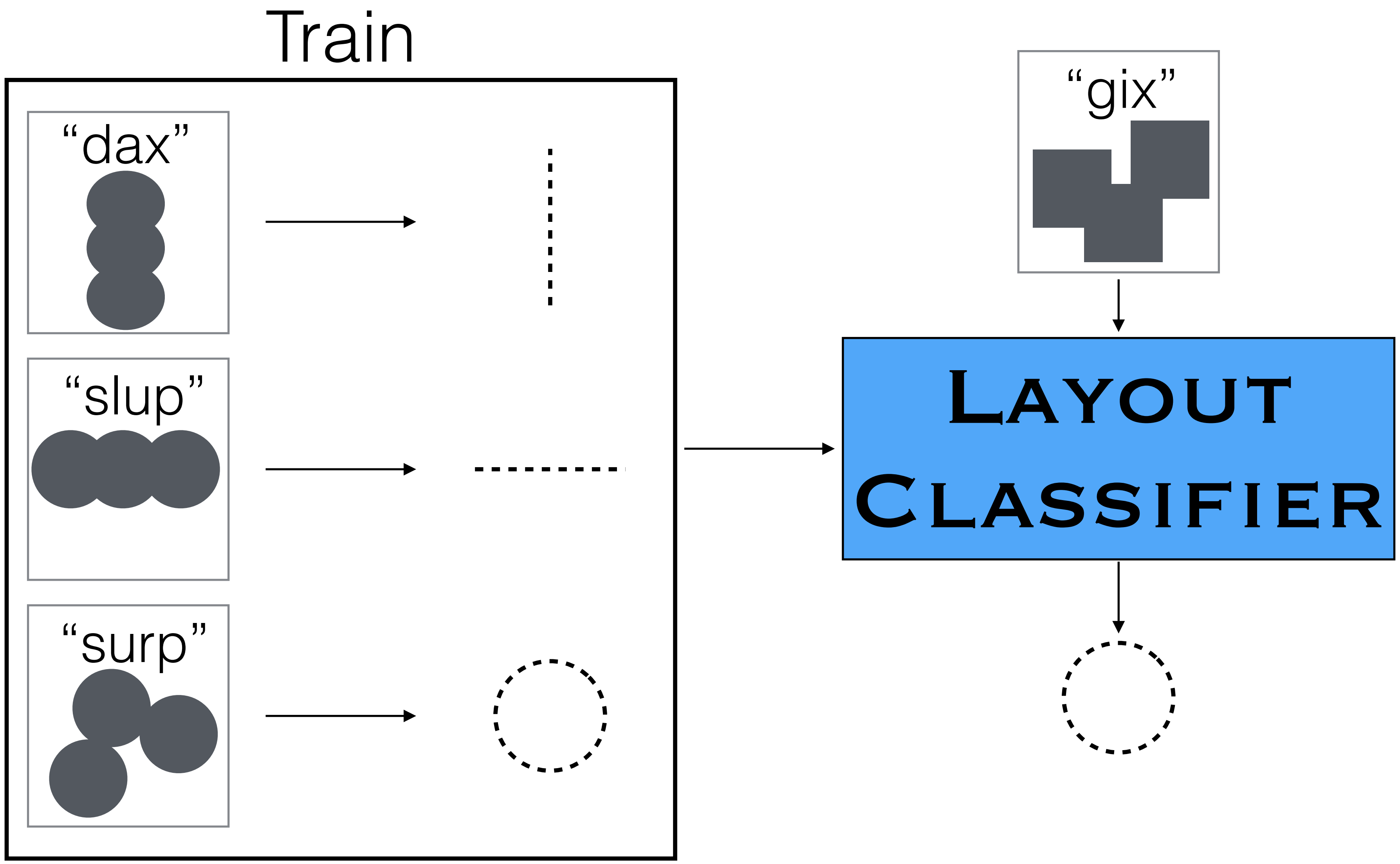
Requirement #2: Concepts represent types



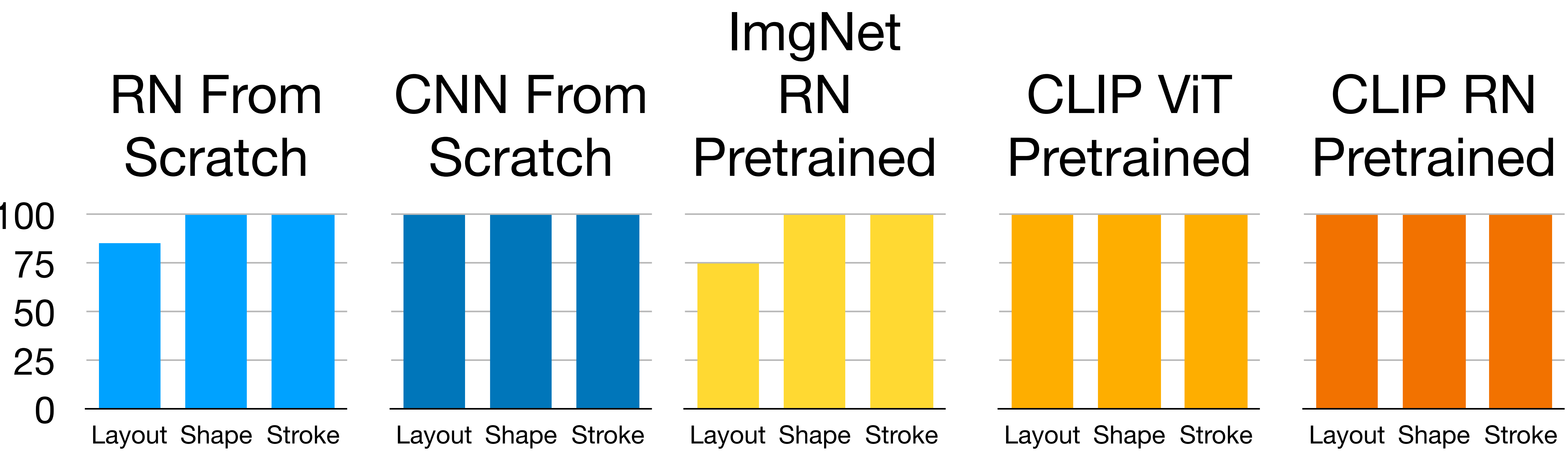
Requirement #2: Concepts represent types



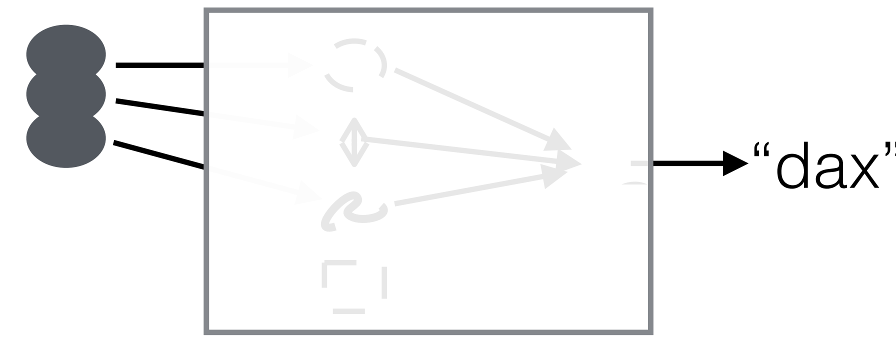
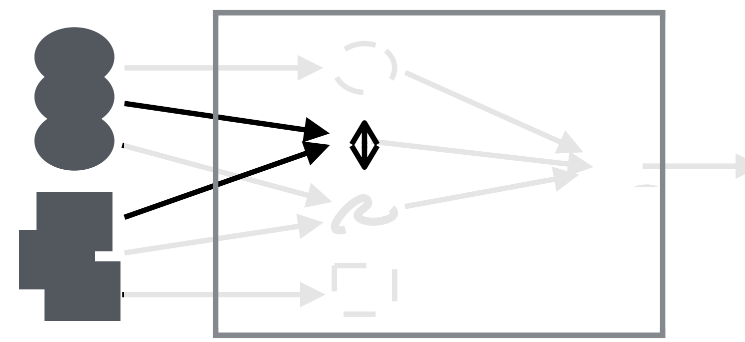
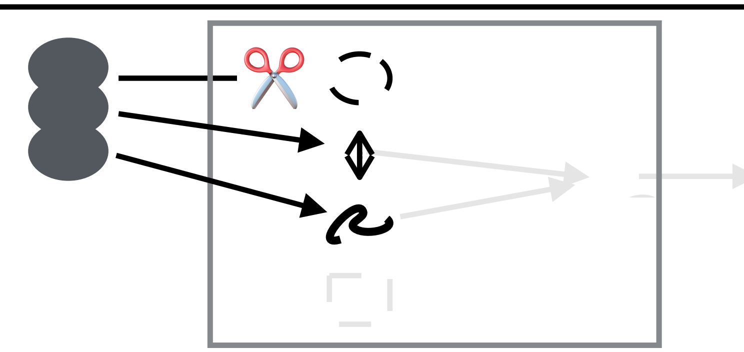
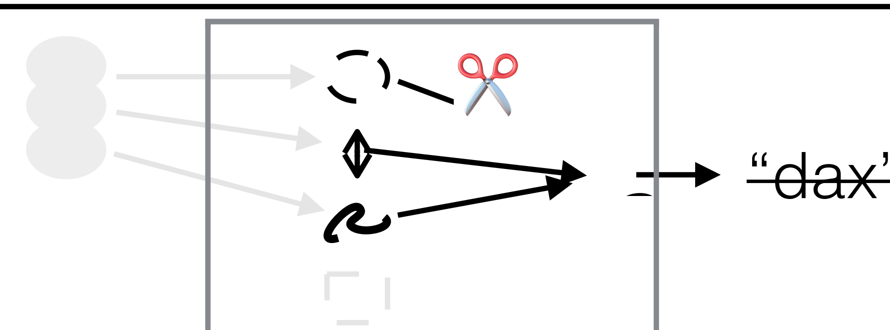
Requirement #2: Concepts represent types



Requirement #2: Concepts represent types



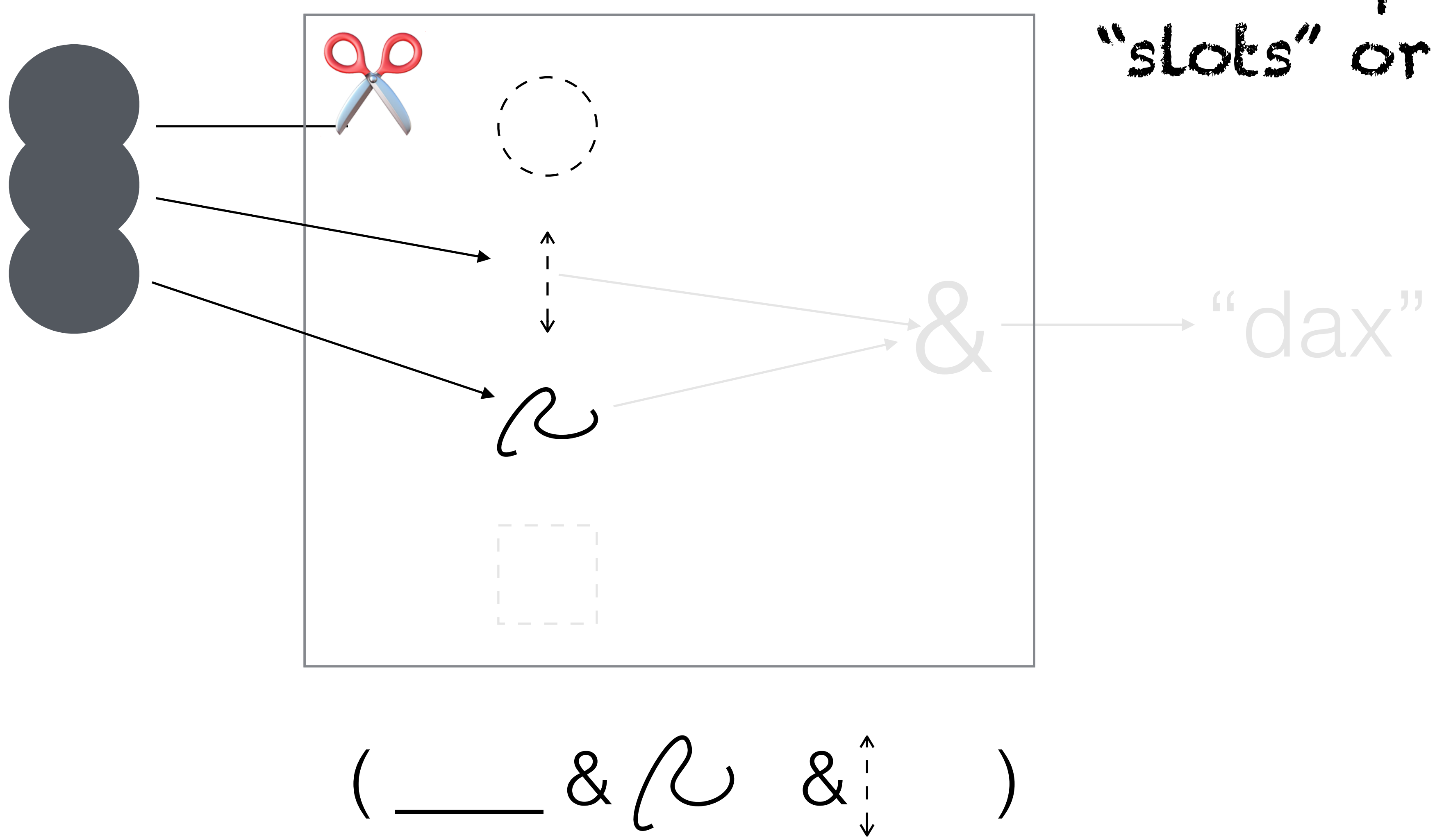
High-Level API

✅	Predictions are **grounded**	
✅	Concepts represent **types**	
	Concepts are **modular**	
	Concepts are **causal**	

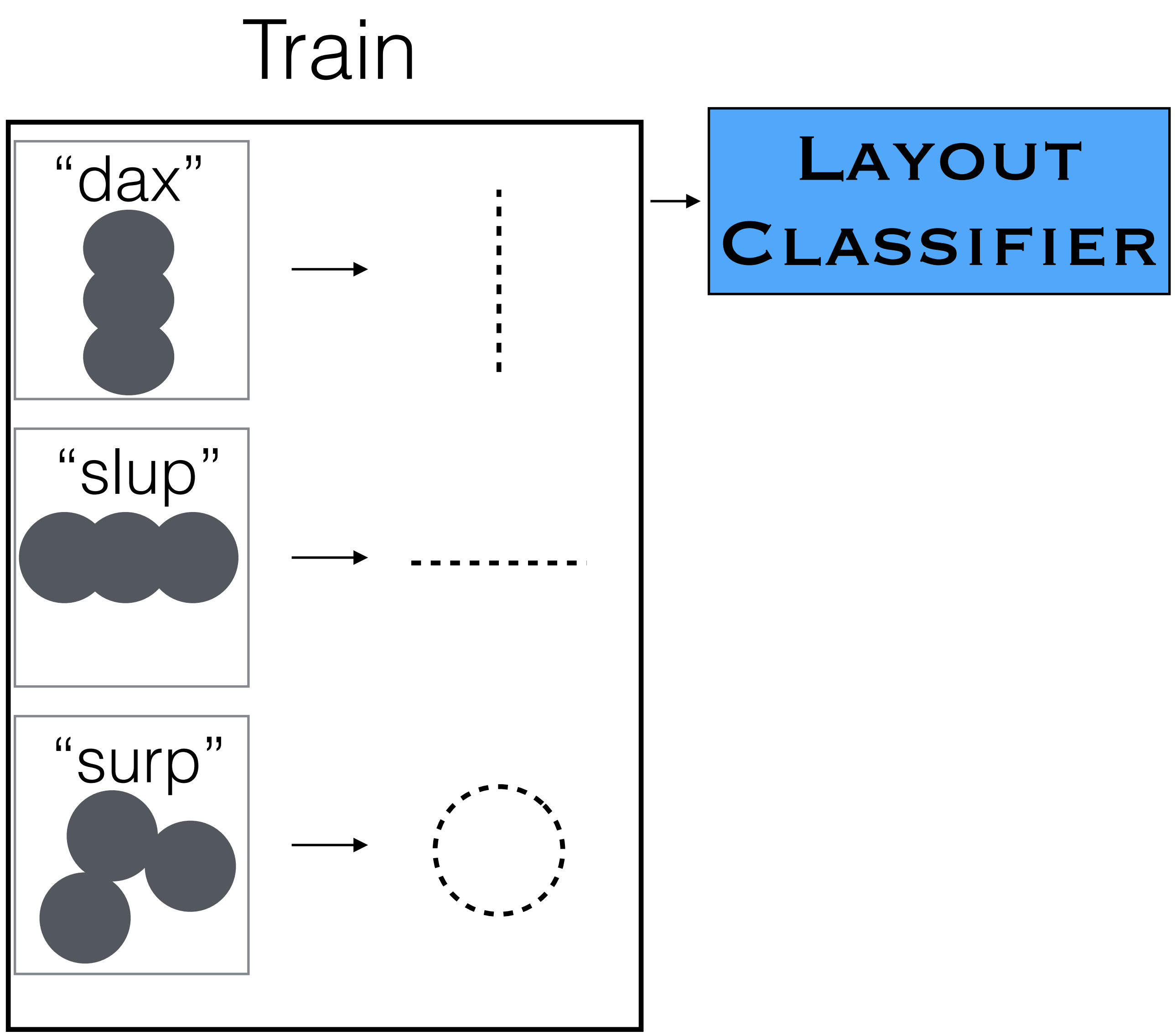
Requirement #3:
Concepts are modular

Requirement #3: Concepts are modular

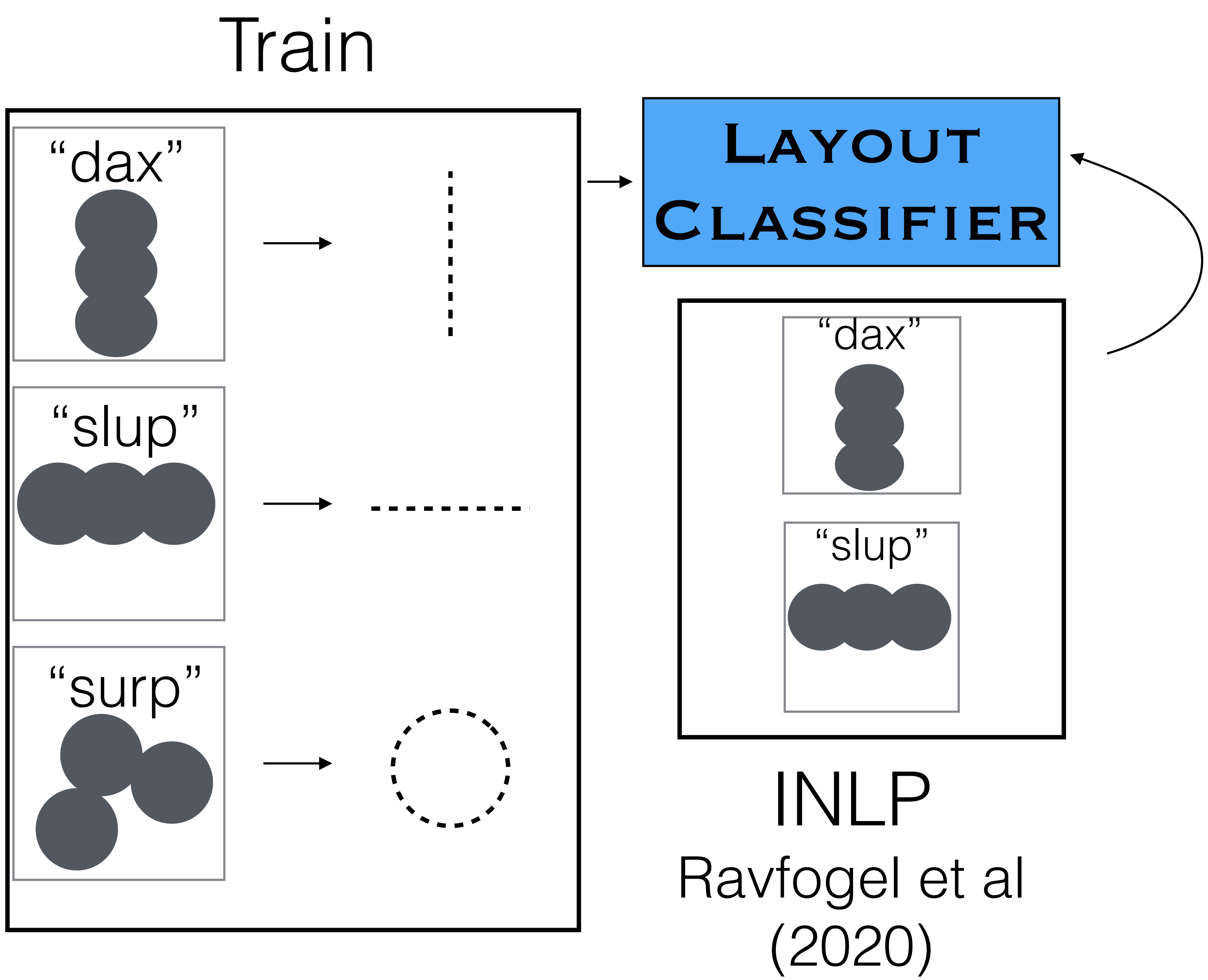
Representation allows decoupling of "slots" or "roles".



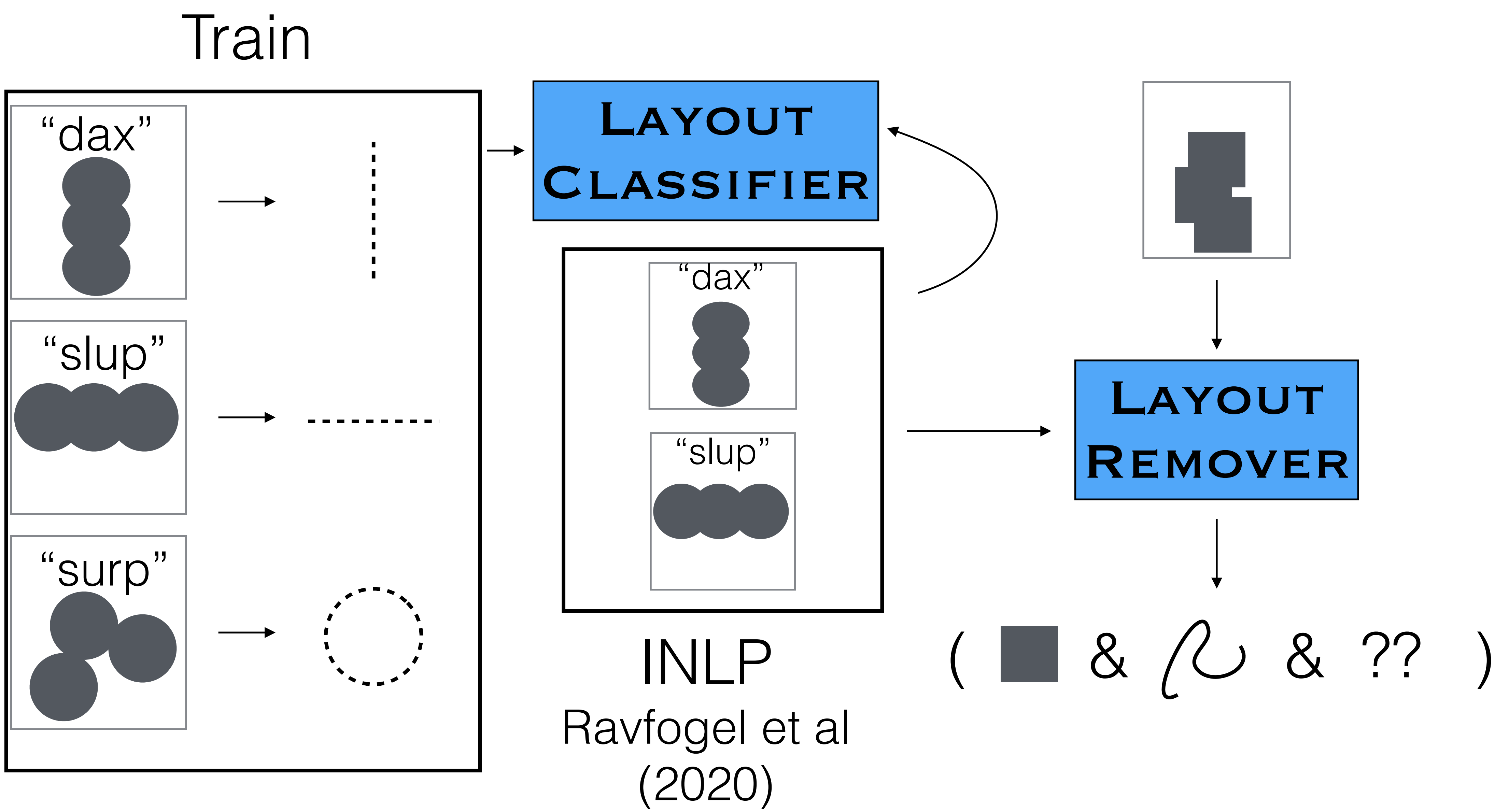
Requirement #3: Concepts are modular



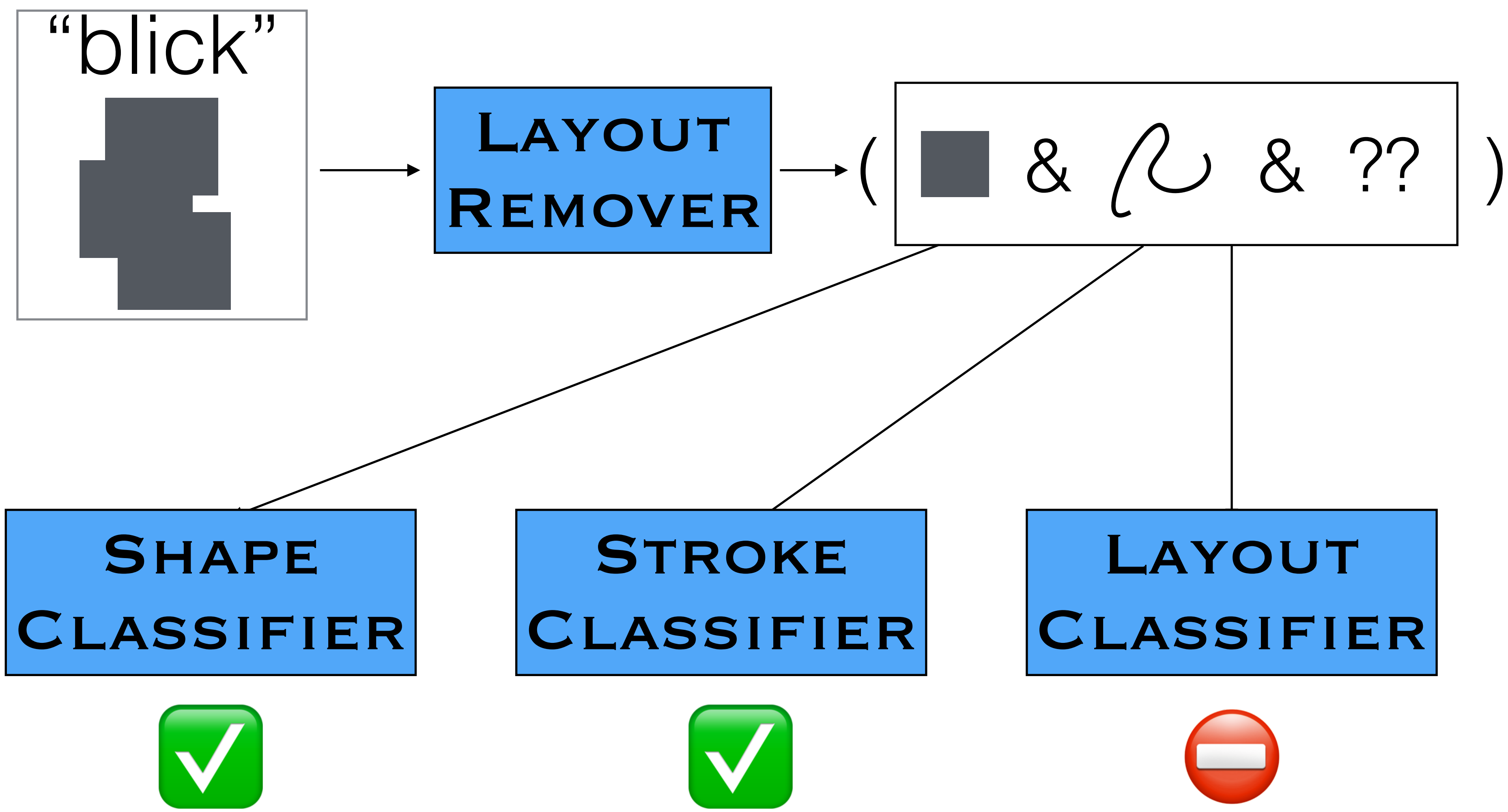
Requirement #3: Concepts are modular



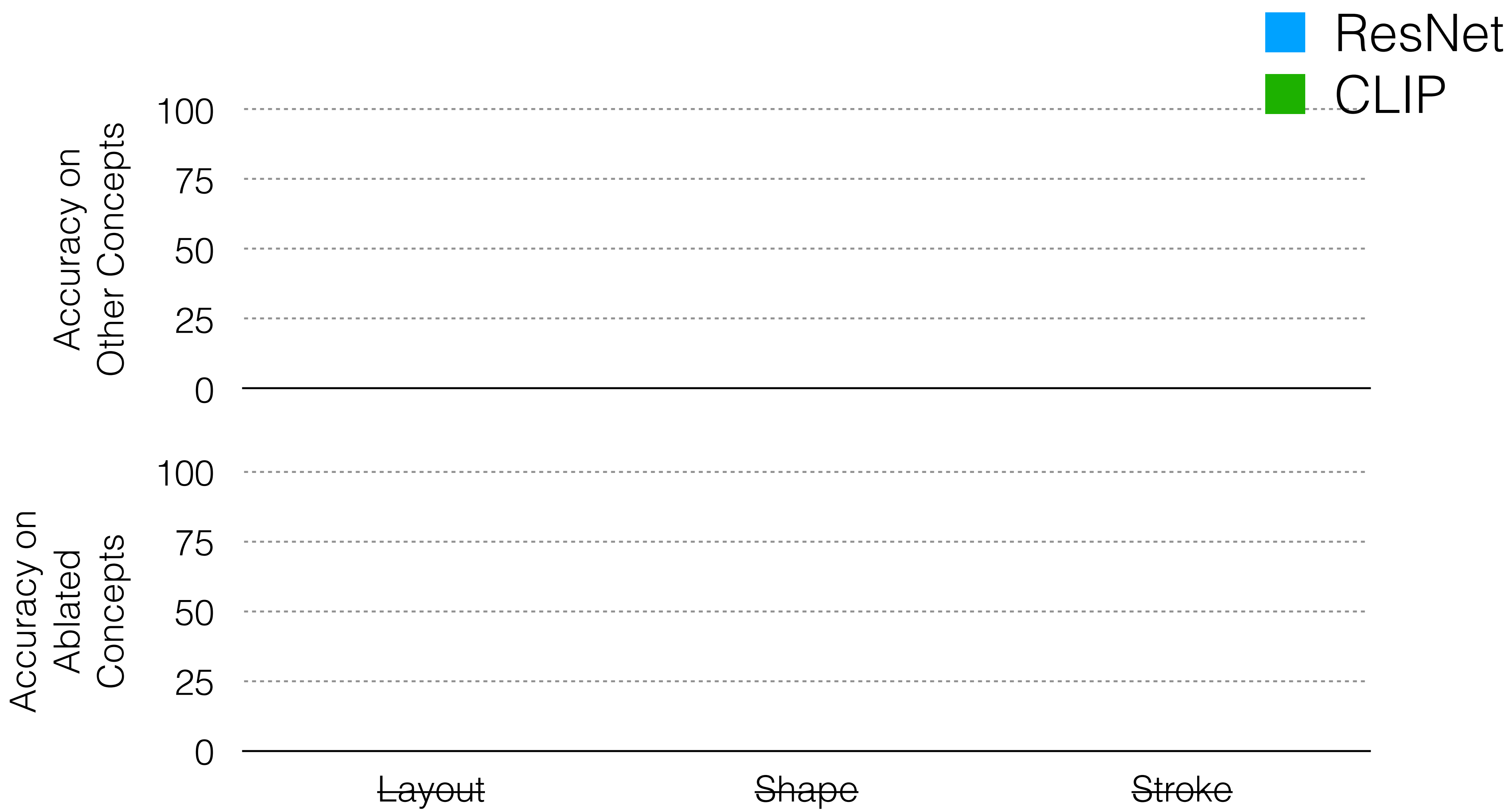
Requirement #3: Concepts are modular



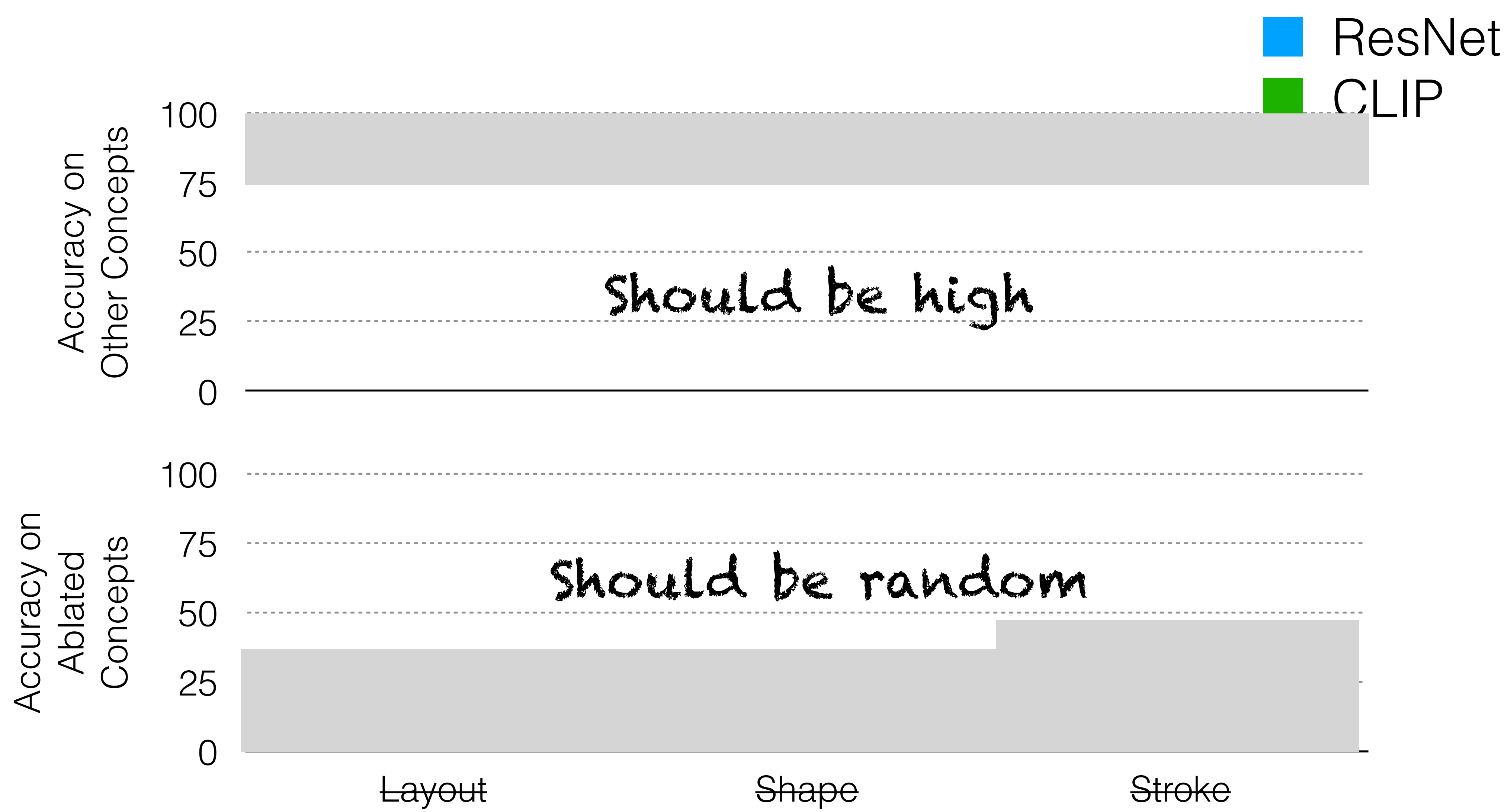
Requirement #3: Concepts are modular



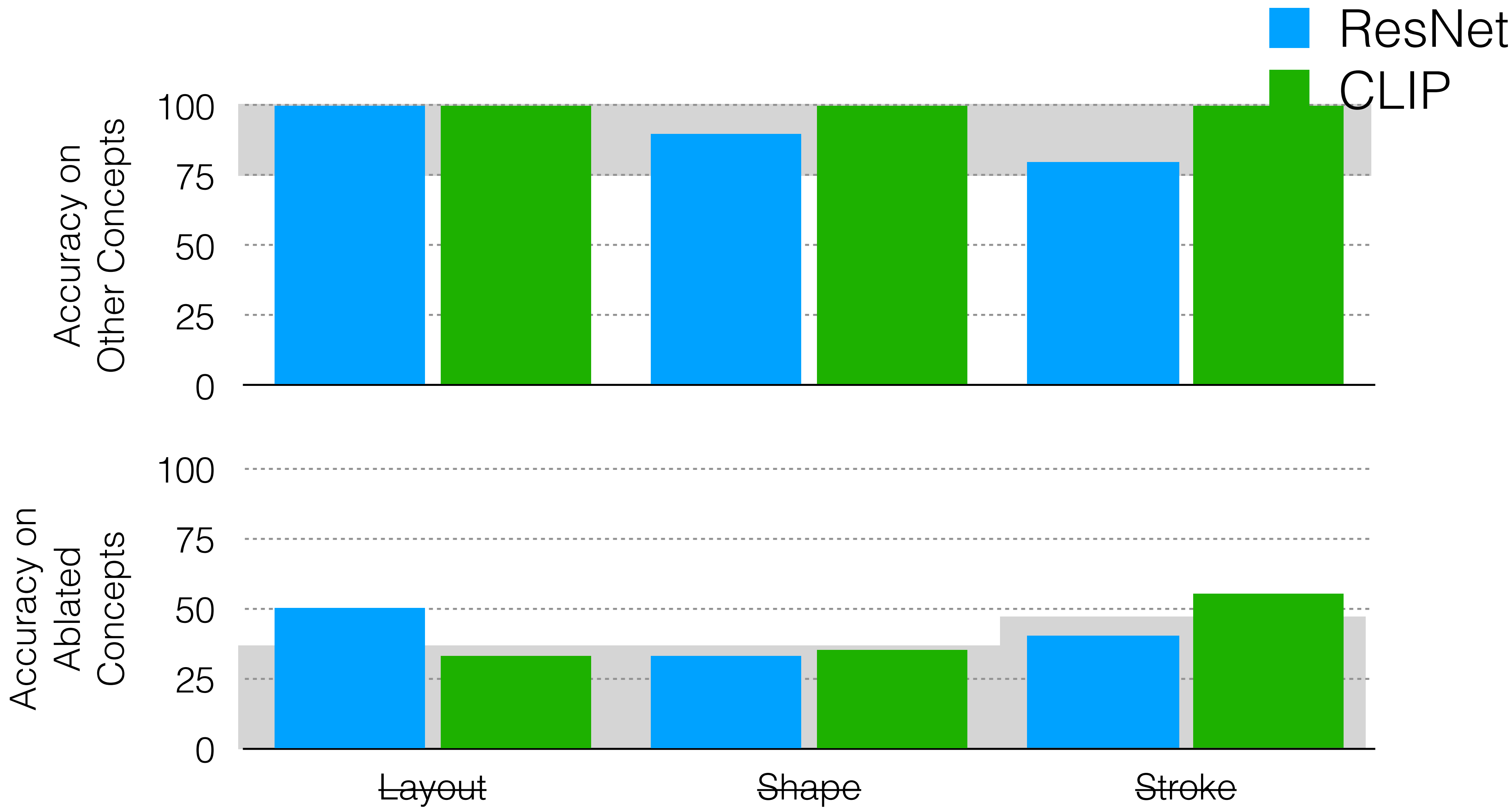
Requirement #3: Concepts are modular



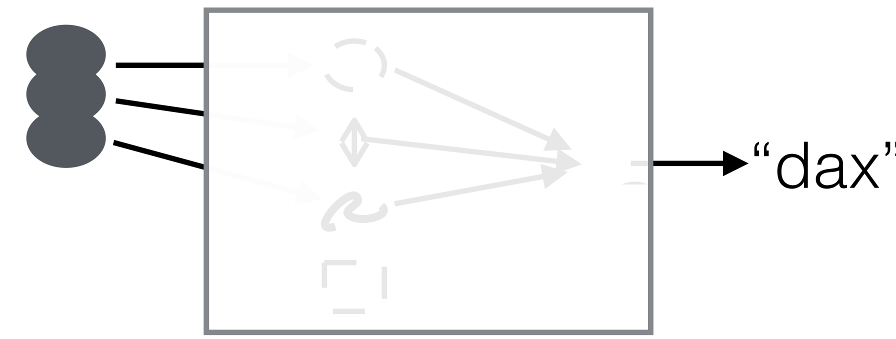
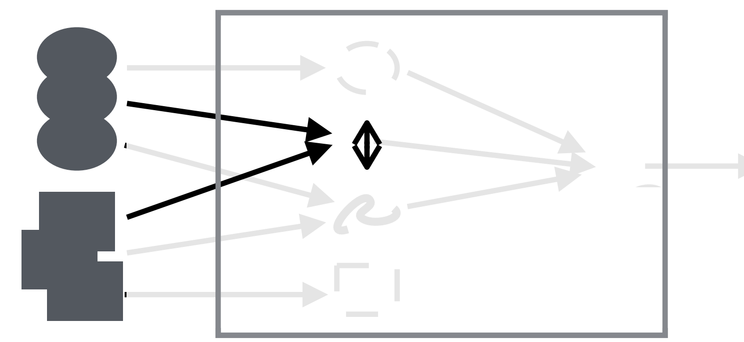
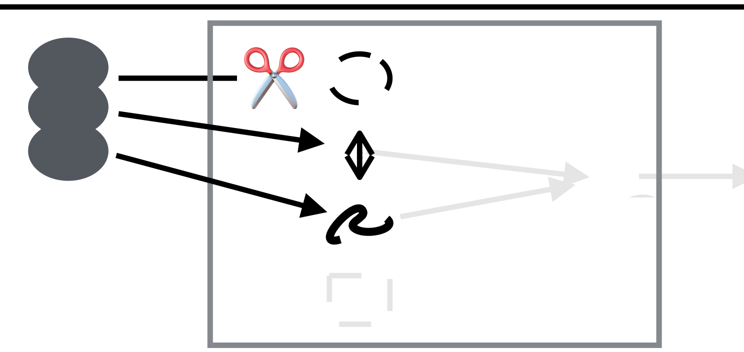
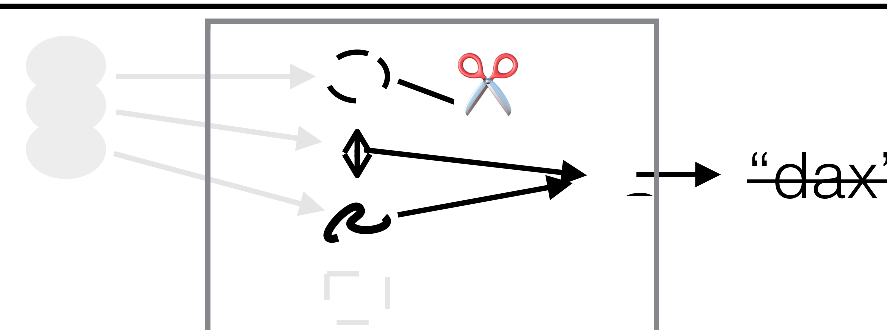
Requirement #3: Concepts are modular



Requirement #3: Concepts are modular

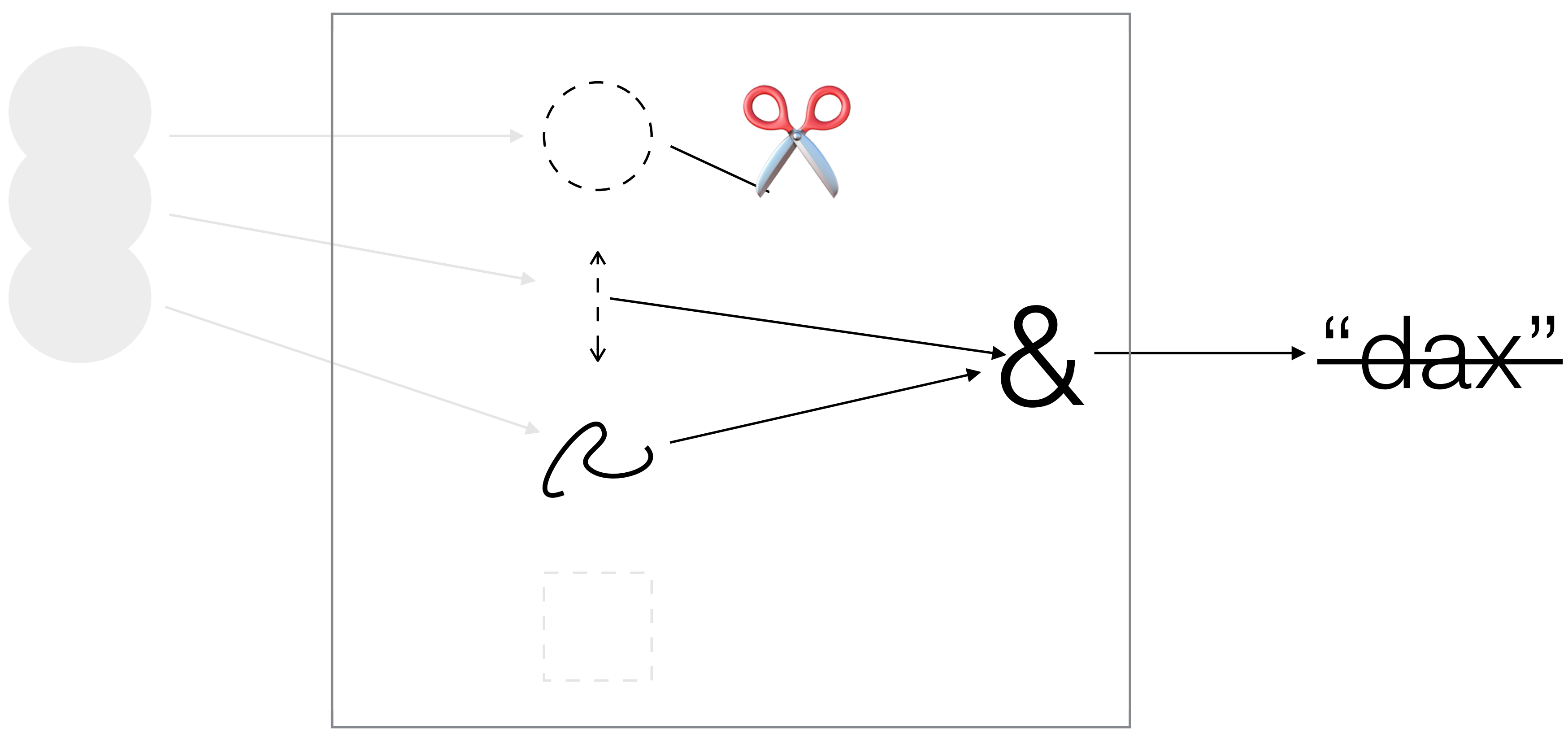


High-Level API

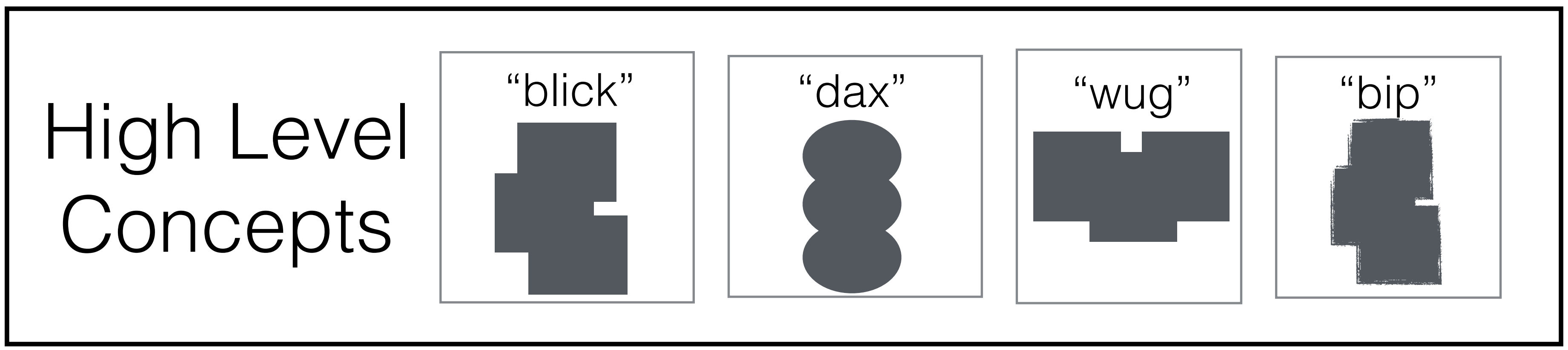
✅	Predictions are **grounded**	
✅	Concepts represent **types**	
✅	Concepts are **modular**	
	Concepts are **causal**	

Requirement #4: Concepts are causal

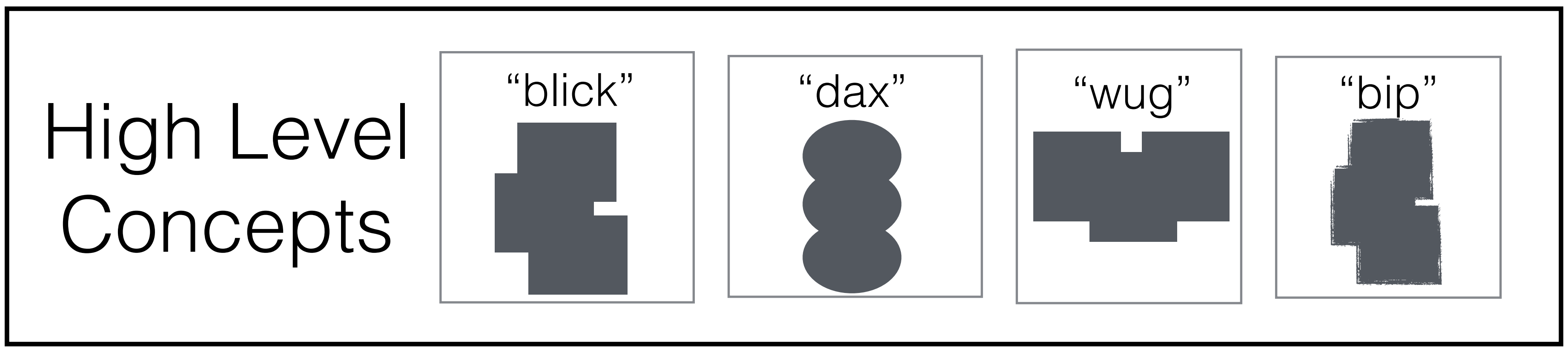
Representations of the parts are causally implicated in the representation of the whole.



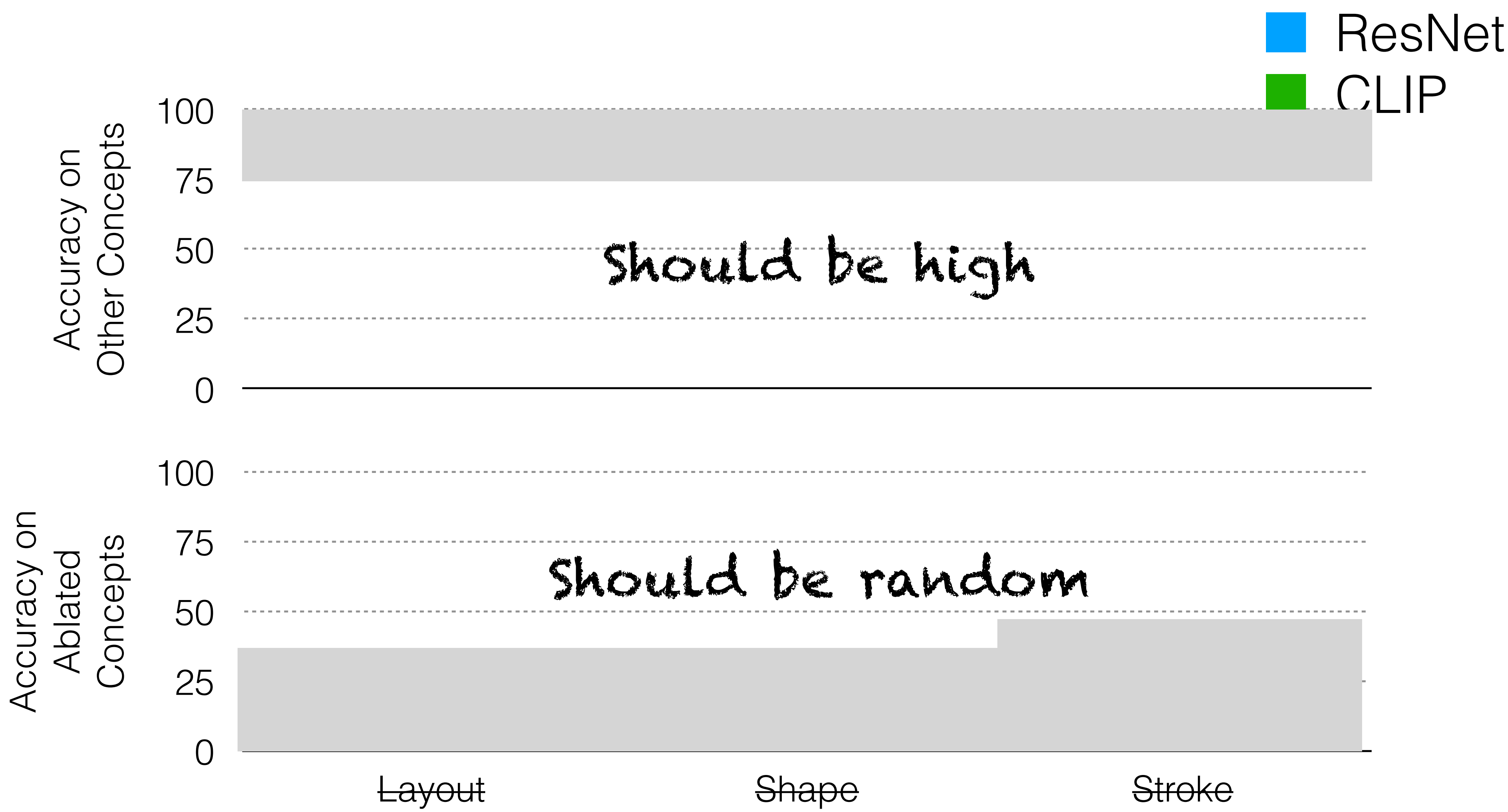
Requirement #4:
Concepts are causal



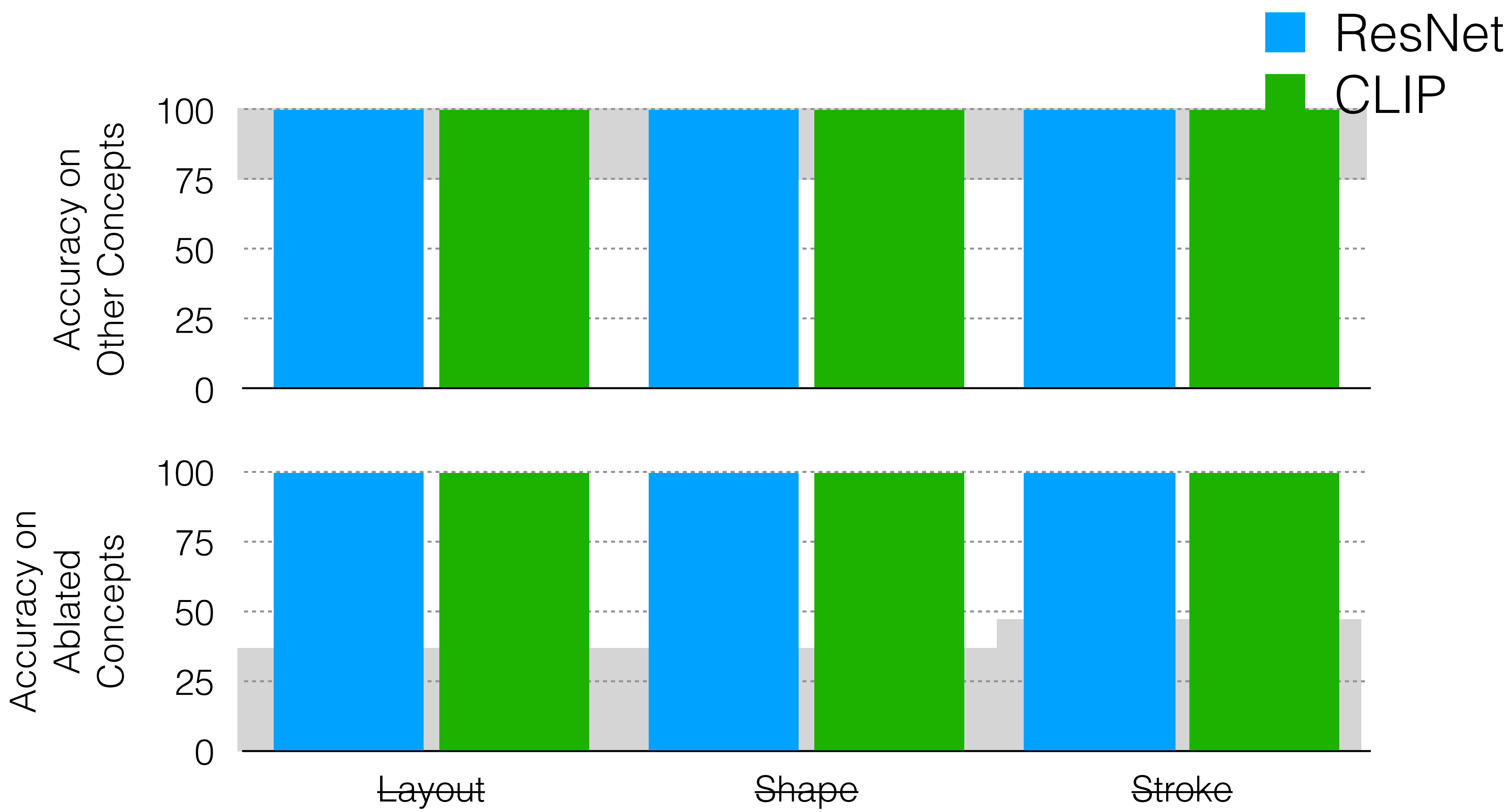
Requirement #4: Concepts are causal



Requirement #4: Concepts are causal



Requirement #4: Concepts are causal

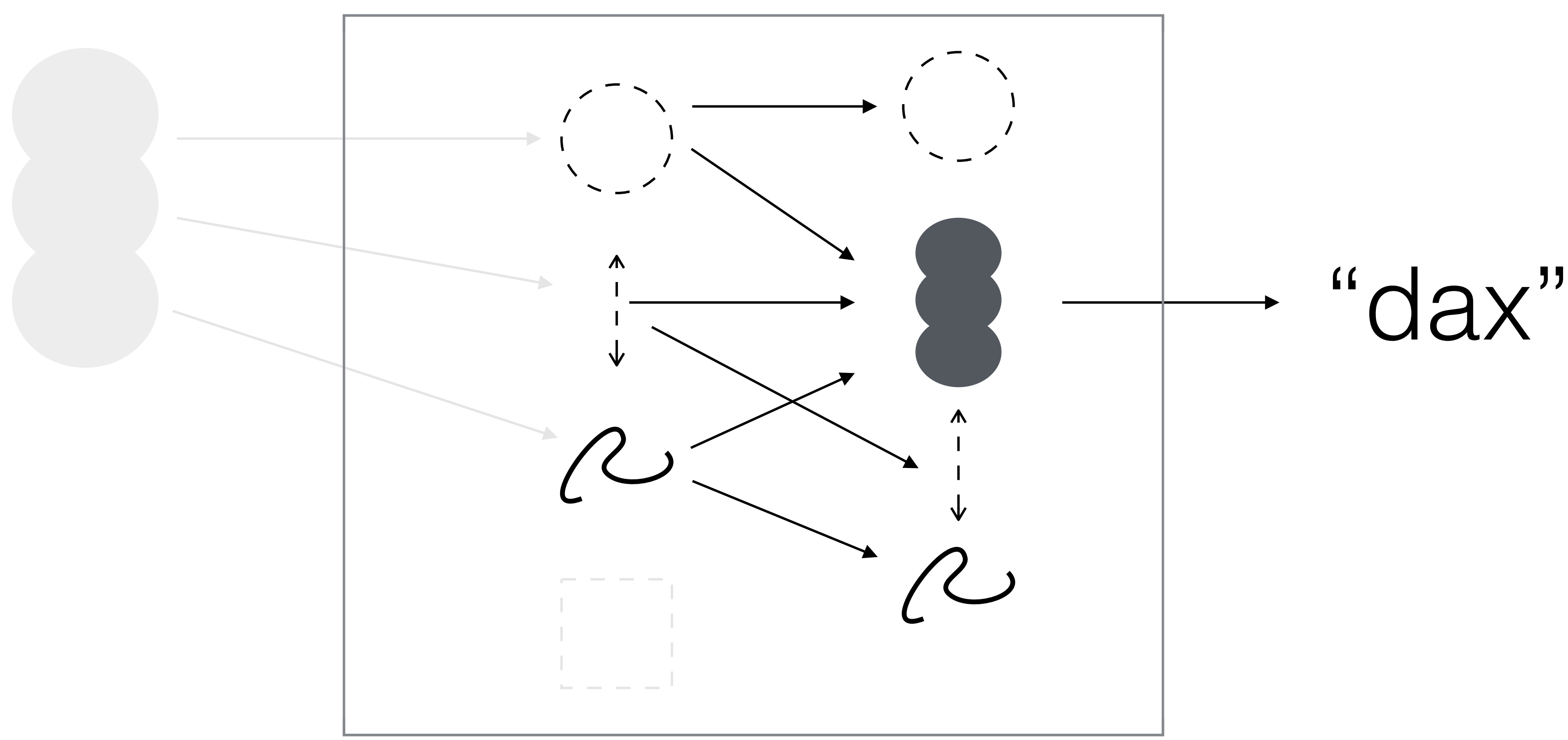


Requirement #4: Concepts are causal

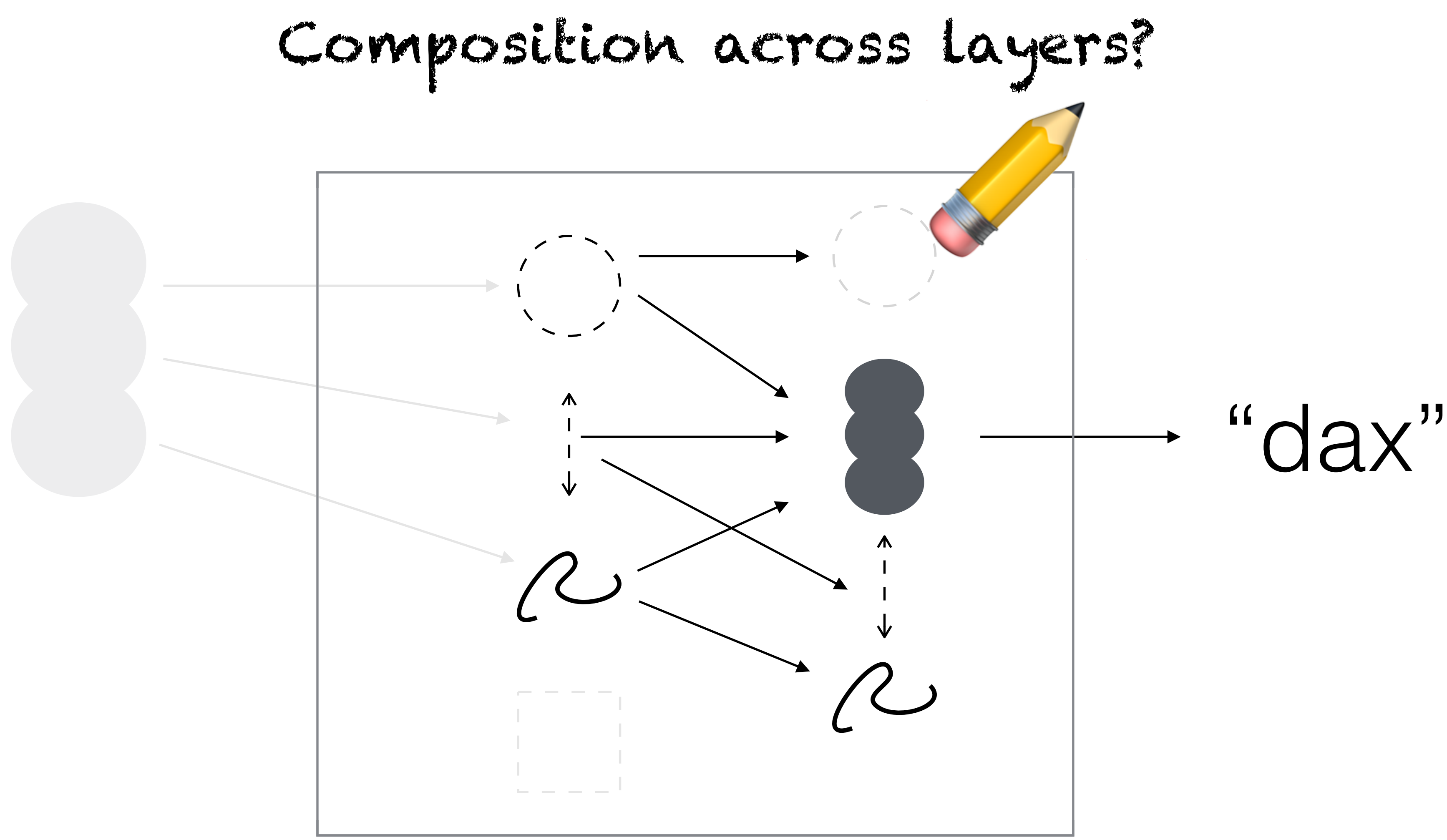
Composition across layers?

Requirement #4: Concepts are causal

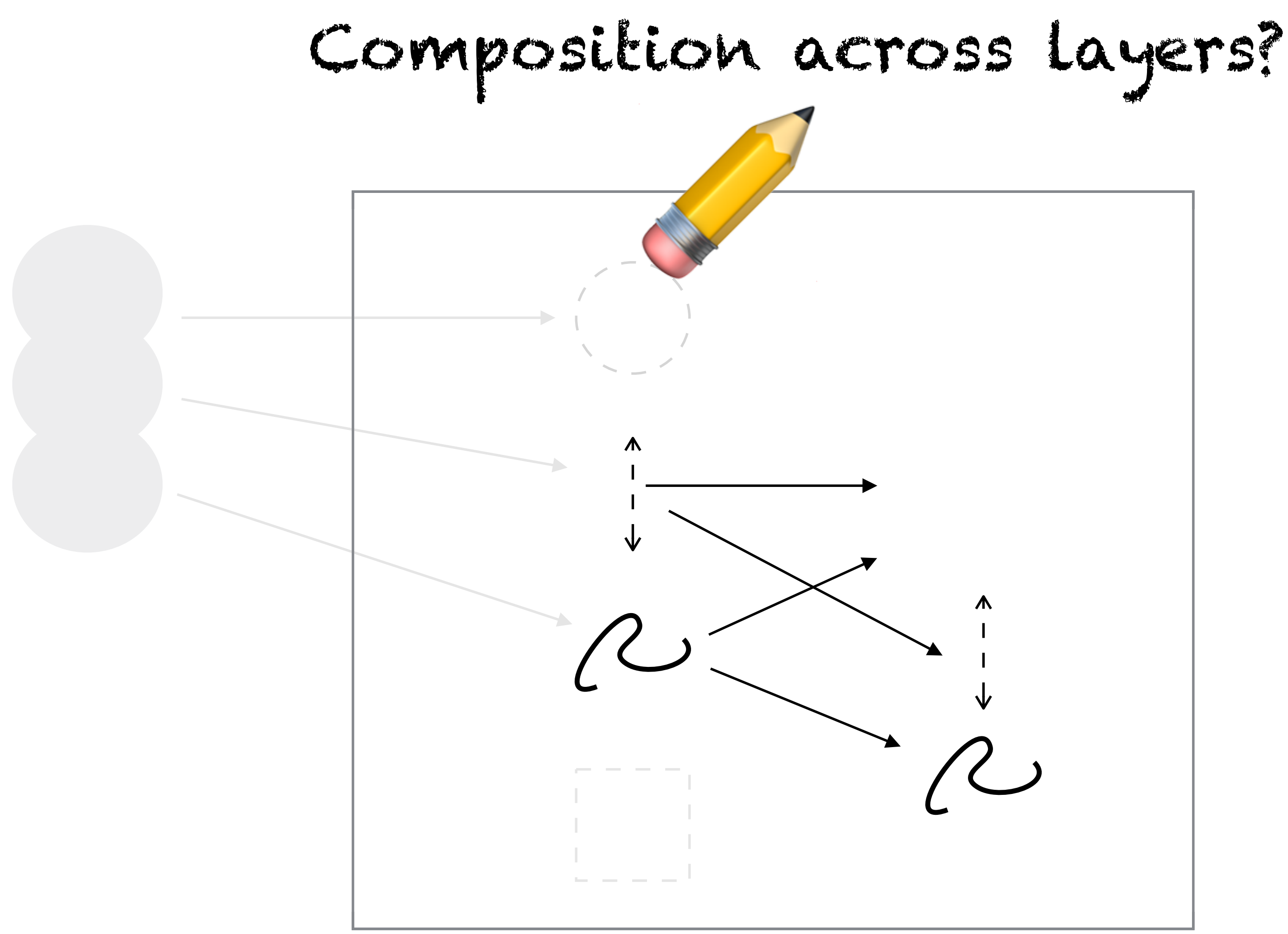
Composition across layers?



Requirement #4: Concepts are causal



Requirement #4: Concepts are causal



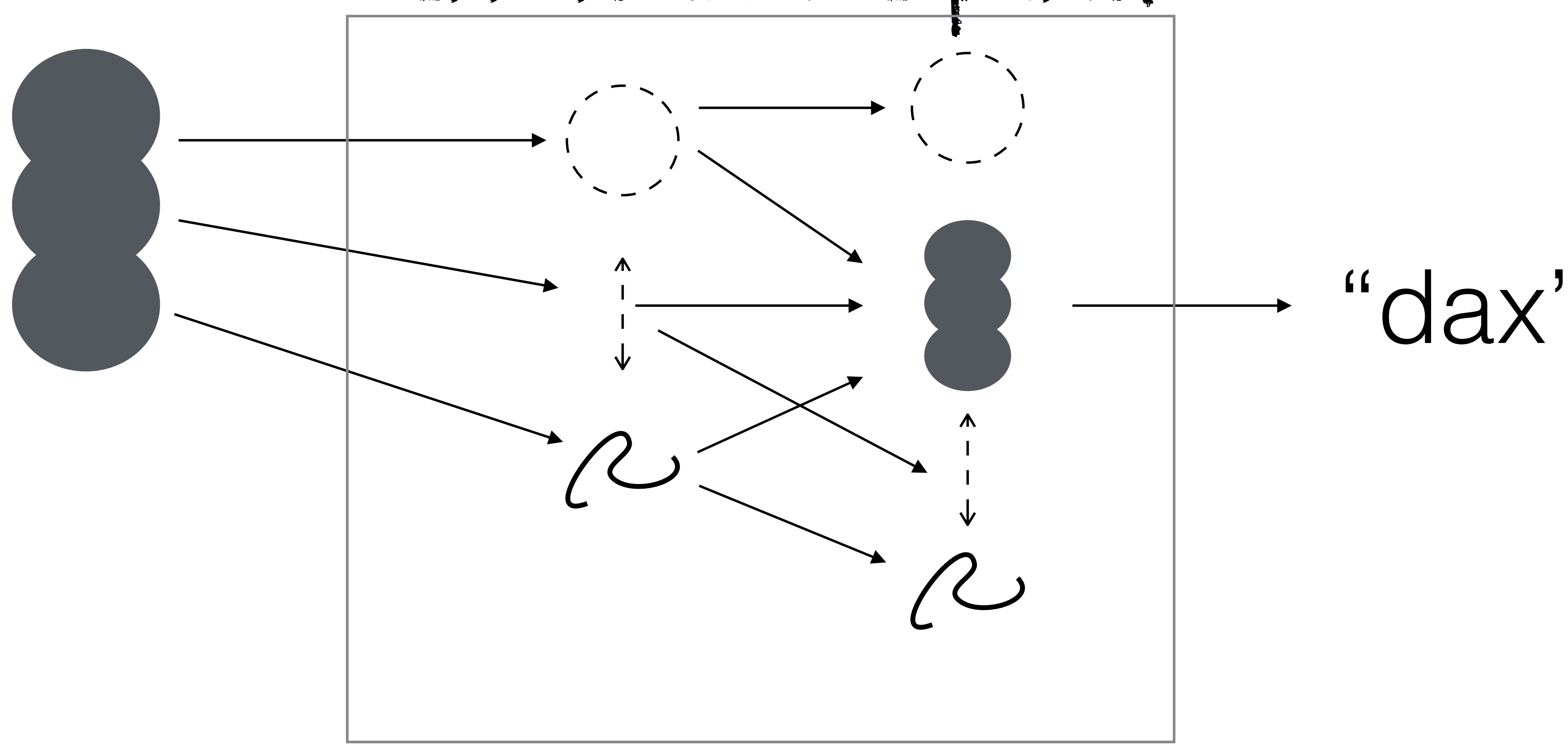
~~"dax"~~

Requirement #4: Concepts are causal

Can errors in the whole be explained by errors in the parts?

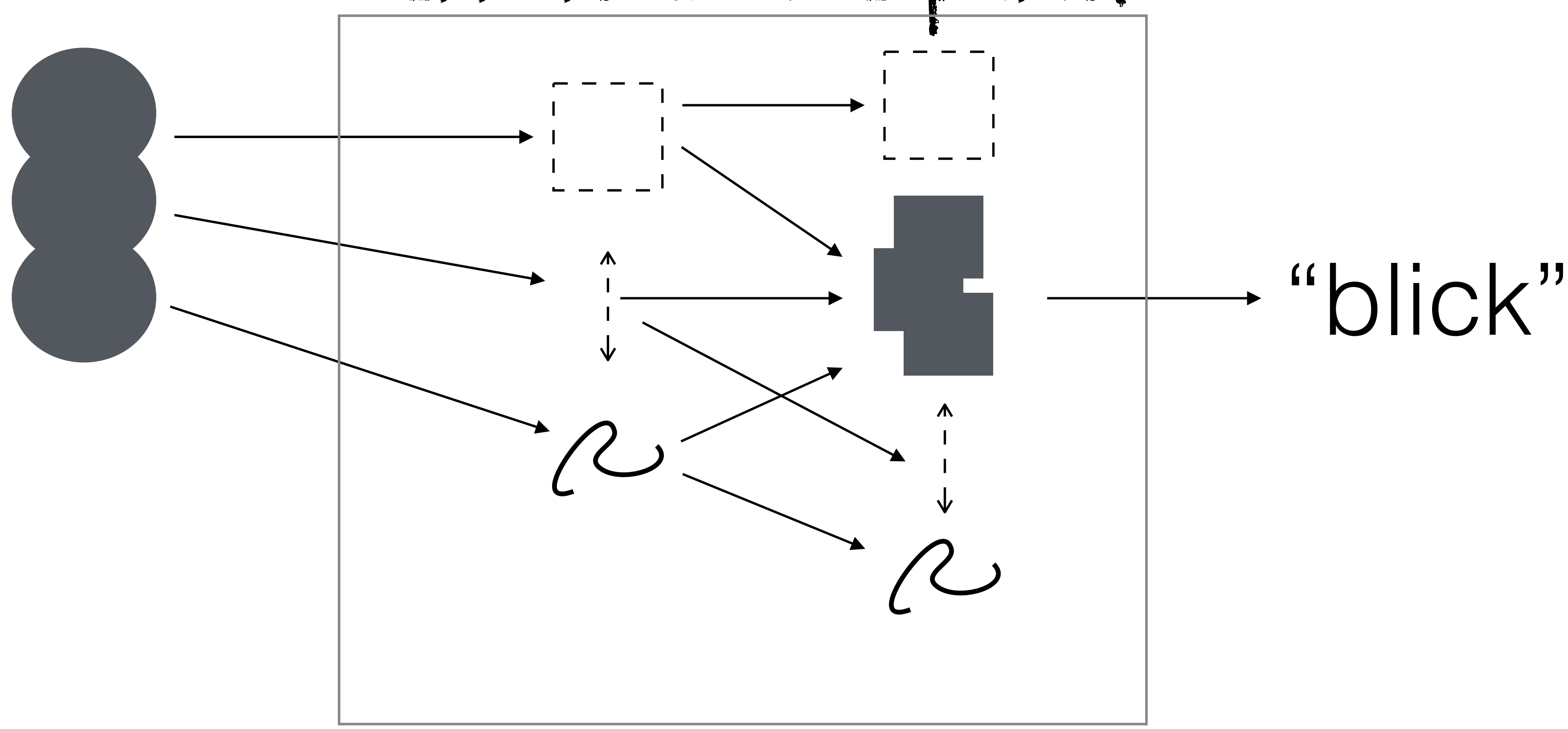
Requirement #4: Concepts are causal

Can errors in the whole be explained by errors in the parts?



Requirement #4: Concepts are causal

Can errors in the whole be explained by errors in the parts?

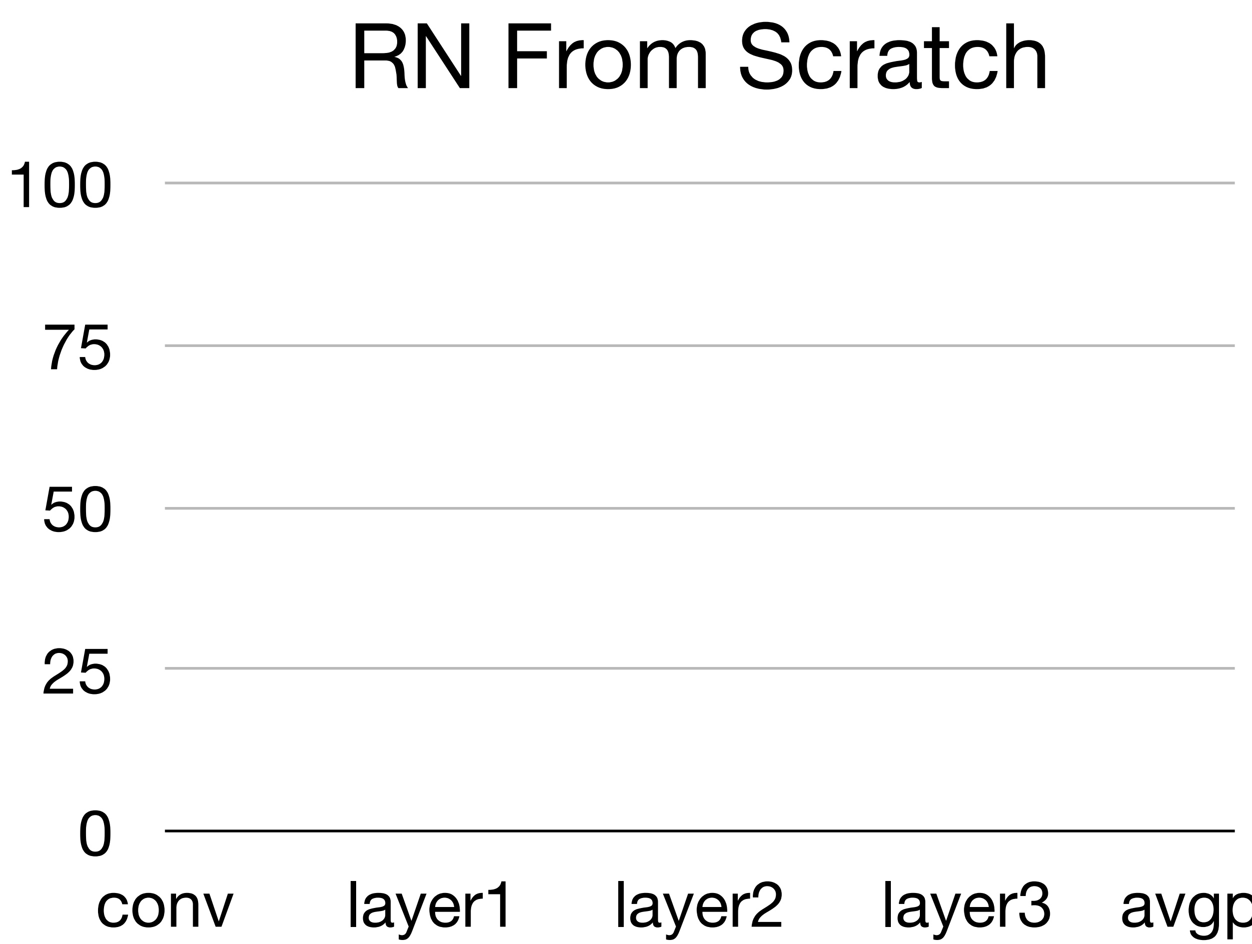


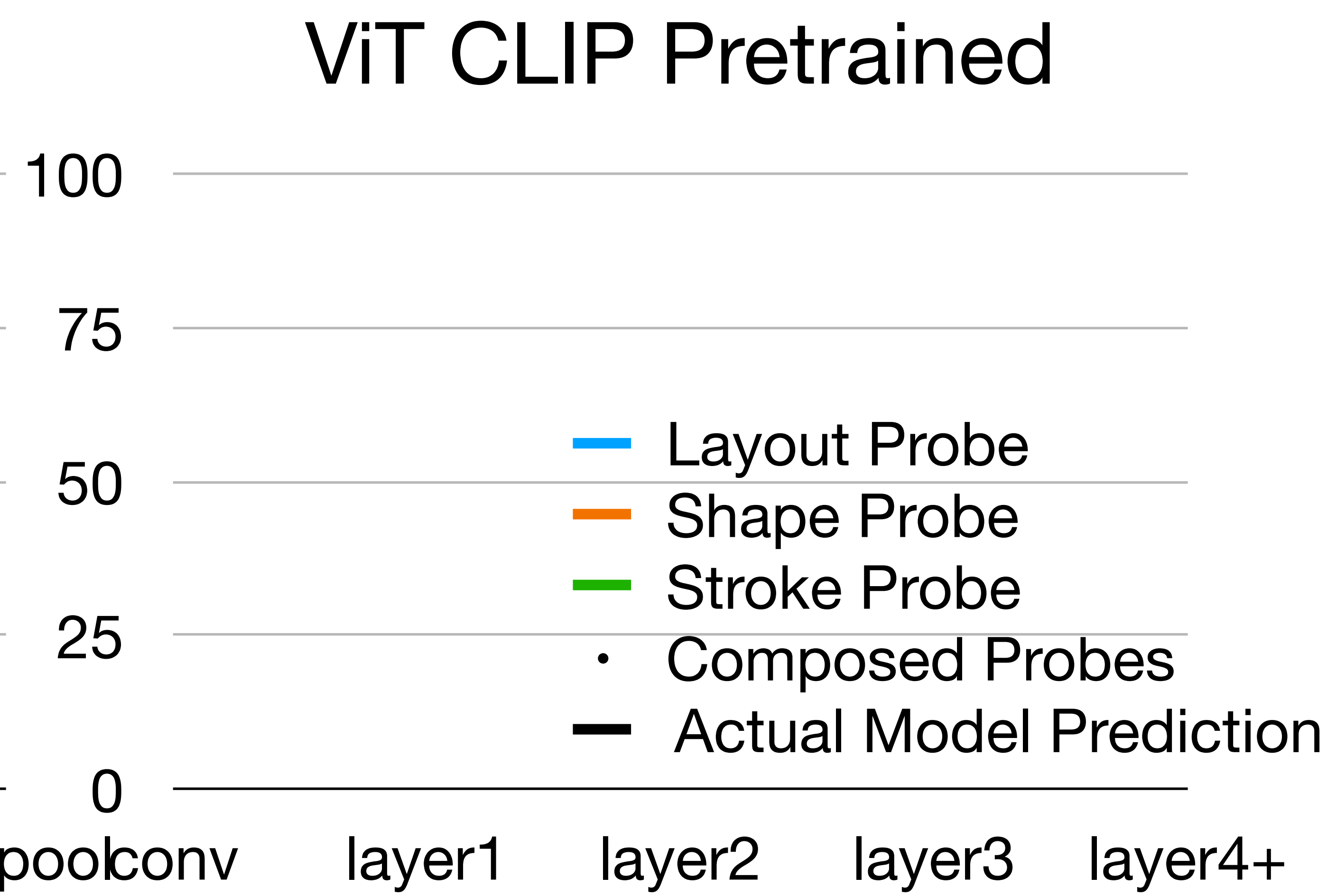
Requirement #4: Concepts are causal

Can errors in the whole be explained by errors in the parts in aggregate?

Requirement #4: Concepts are causal

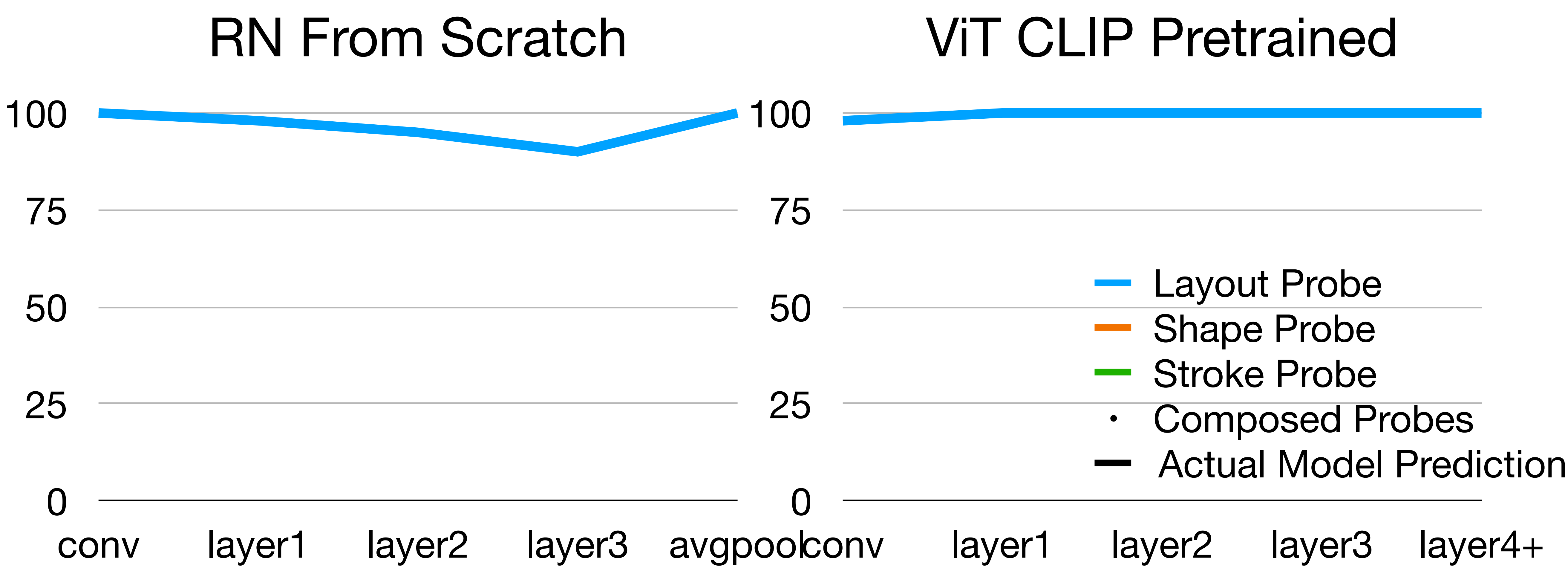
Can errors in the whole be explained by errors in the parts in aggregate?





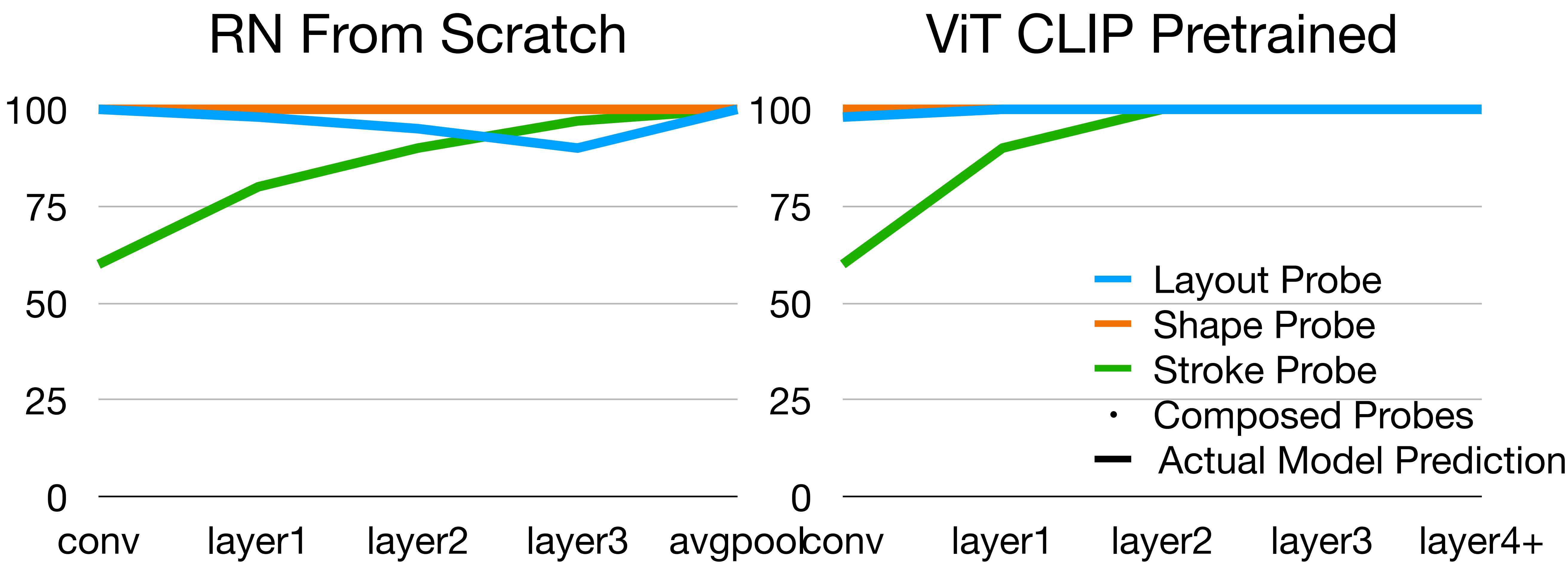
Requirement #4: Concepts are causal

Can errors in the whole be explained by errors in the parts in aggregate?



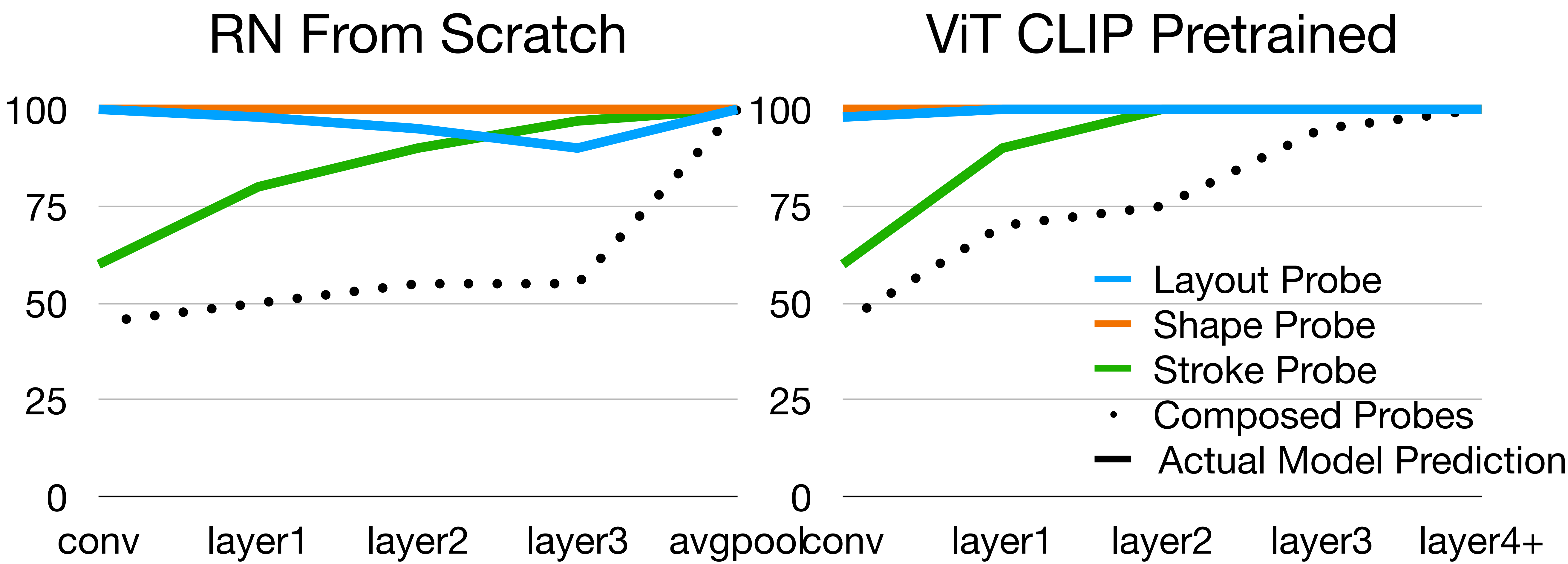
Requirement #4: Concepts are causal

Can errors in the whole be explained by errors in the parts in aggregate?



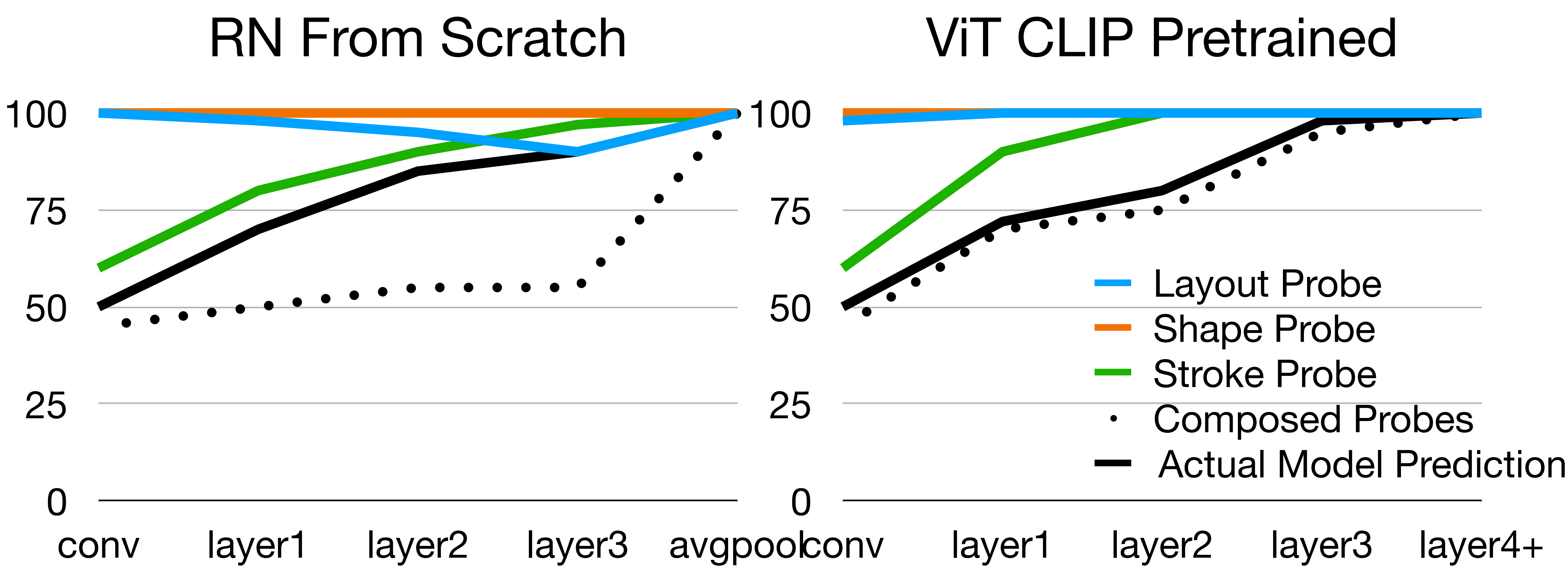
Requirement #4: Concepts are causal

Can errors in the whole be explained by errors in the parts in aggregate?



Requirement #4: Concepts are causal

Can errors in the whole be explained by errors in the parts in aggregate?

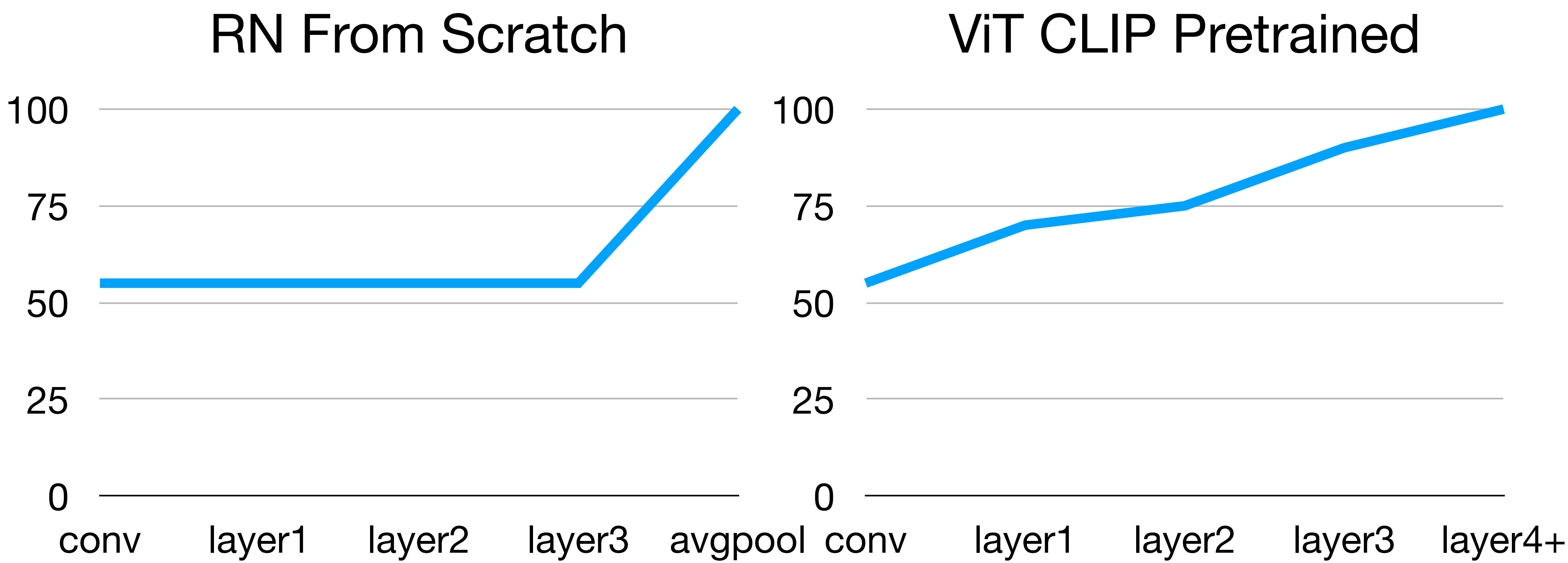


Requirement #4:
Concepts are causal

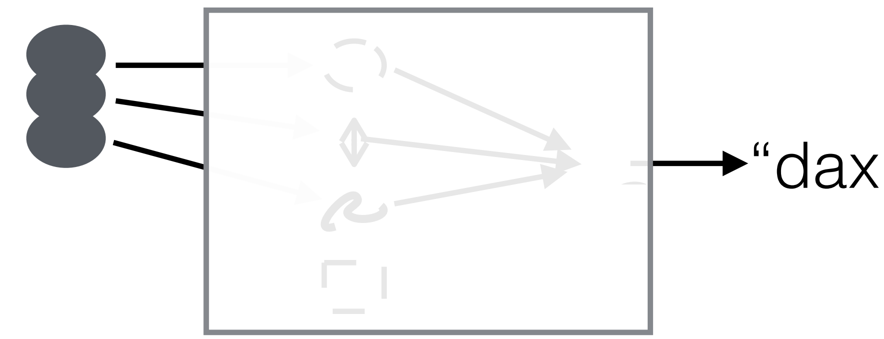
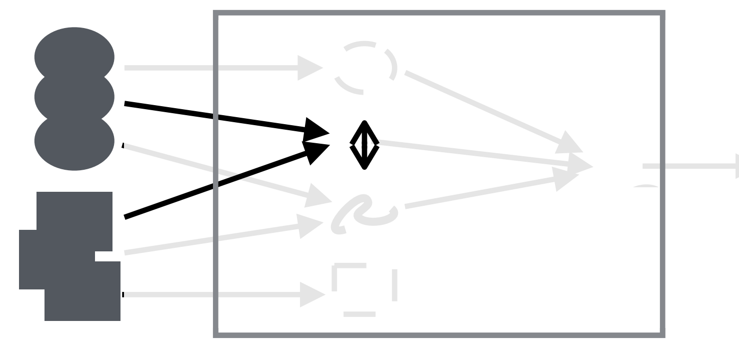
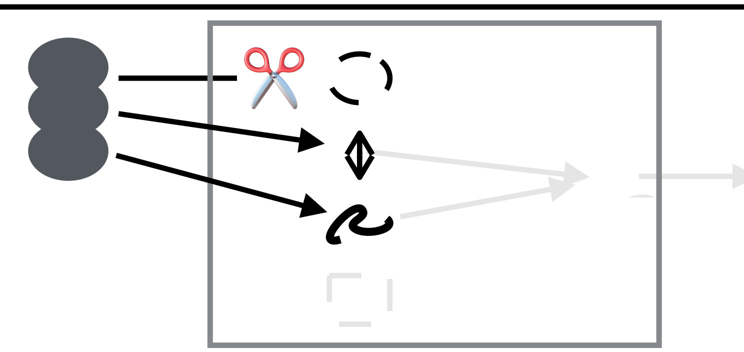
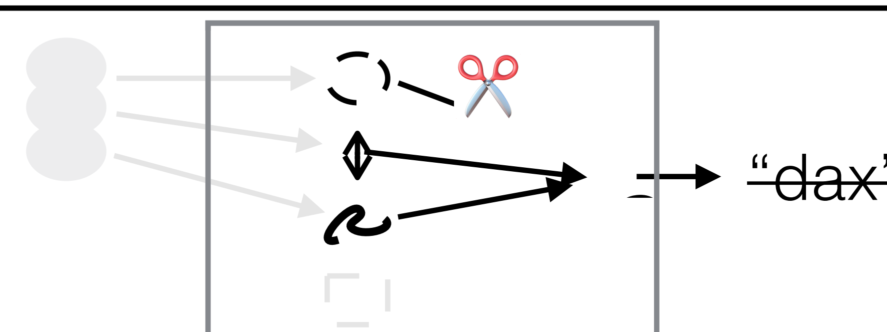
Can errors in the whole be explained by errors in the parts at the instance level?

Requirement #4: Concepts are causal

Can errors in the whole be explained by errors in the parts at the instance level?



High-Level API

✅	Predictions are **grounded**	
✅	Concepts represent **types**	
✅	Concepts are **modular**	
⚠️	Concepts are **causal**	

Takeaways

- When learning to discriminate visual concepts, end-to-end NNs learn complex internal representations
- These representations meet basic criteria of “structured” compositional representations
    - They are grounded in the external world
    - Complex concepts are built from reusable parts
    - Parts are sufficiently disentangled
    - Representations of parts might be causally implicated in representations of wholes
- Pretrained models show some advantage, but results are preliminary
    - Some desirable inductive biases (shape > color in object naming)
    - Pretrained transformer might fair better on causality tests

Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. [TACL 2022]

Jason Wei, Dan Garrette, Tal Linzen and Ellie Pavlick. Frequency Effects on Syntactic Rule Learning in Transformers. [EMNLP 2021]

Charles Lovering, **Rohan Jha**, Tal Linzen and Ellie Pavlick. Predicting Inductive Biases of Pretrained Models. [ICLR 2021]

Aaron Traylor, Roman Feiman and Ellie Pavlick. AND does not mean OR: Using Formal Languages to Study Language Models' Representations. [ACL 2021]

Roma Patel and Ellie Pavlick. Mapping Language Models to Grounded Conceptual Spaces. [ICLR 2022]

Syntactic Concepts and Rules

The dog that chases the cats ___ fast

run

runs

Syntactic Concepts and Rules

The dog that chases the cats runs fast

Syntactic Concepts and Rules

The **dog** that chases the cats **runs** fast

Syntactic Concepts and Rules

The **dog** that chases the cats **runs** fast

SINGULAR NOUN

SINGULAR VERB

Syntactic Concepts and Rules

The **dog** that chases the cats **runs** fast

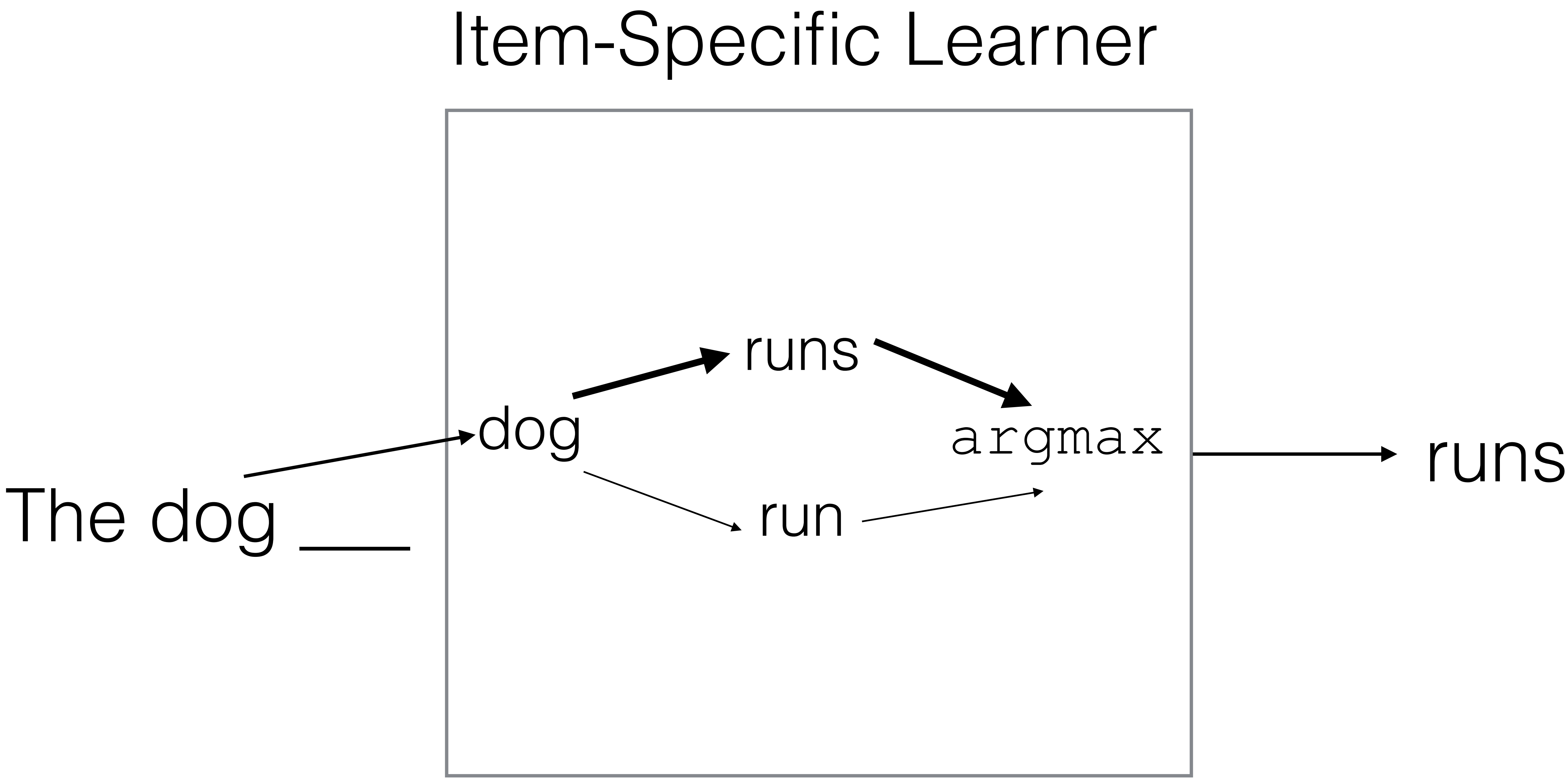
SINGULAR NOUN

SINGULAR VERB

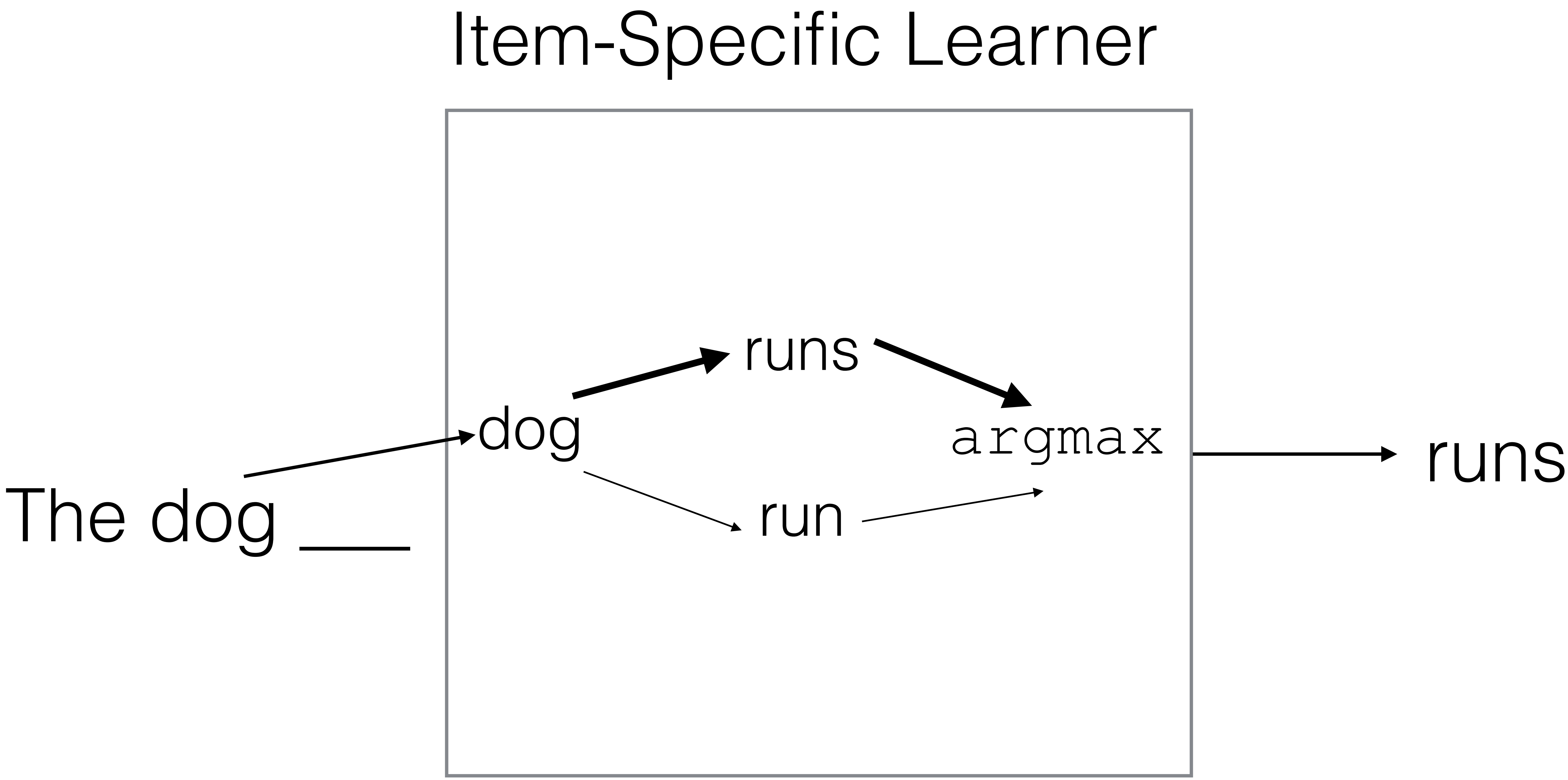
`if` SINGULAR NOUN `then` SINGULAR VERB

Categories of Reasoning

Categories of Reasoning



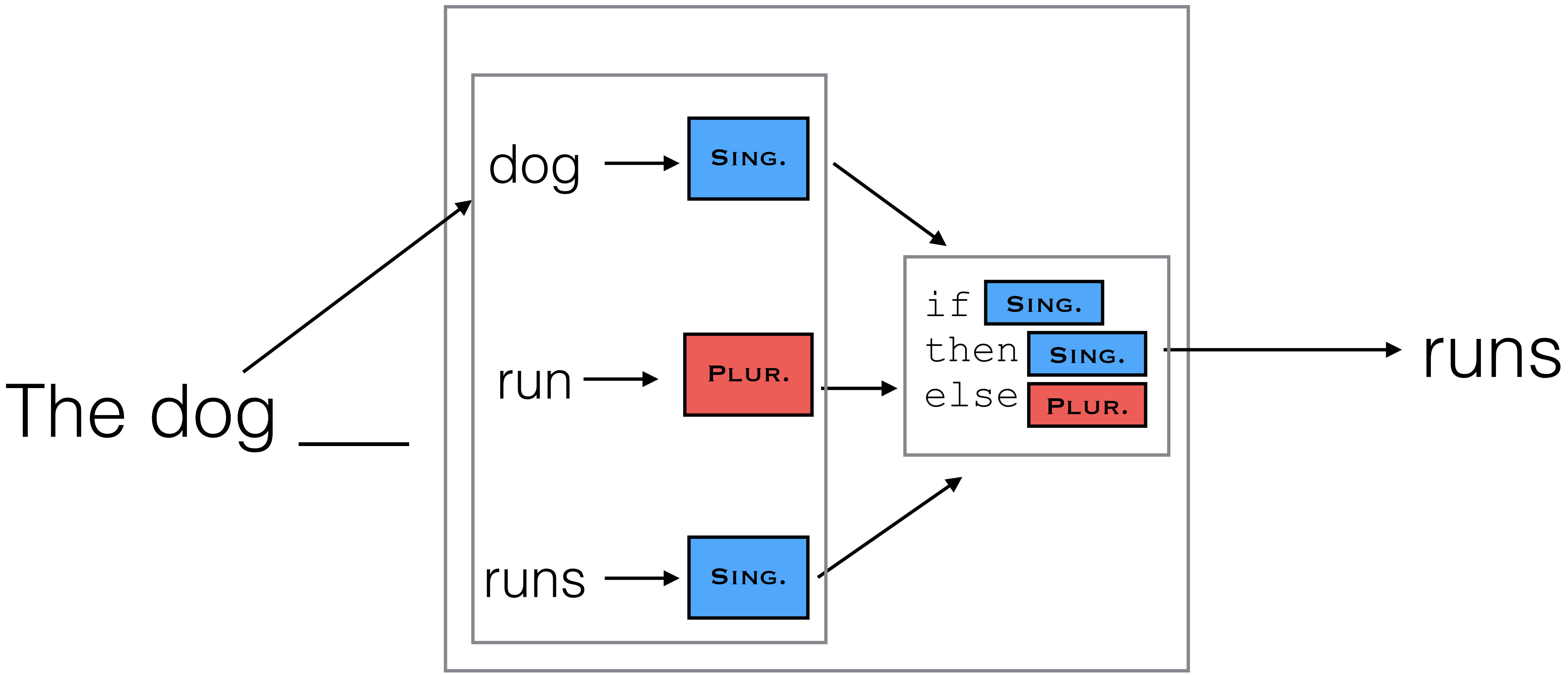
Categories of Reasoning



Decision depends entirely on specifics of inputs.

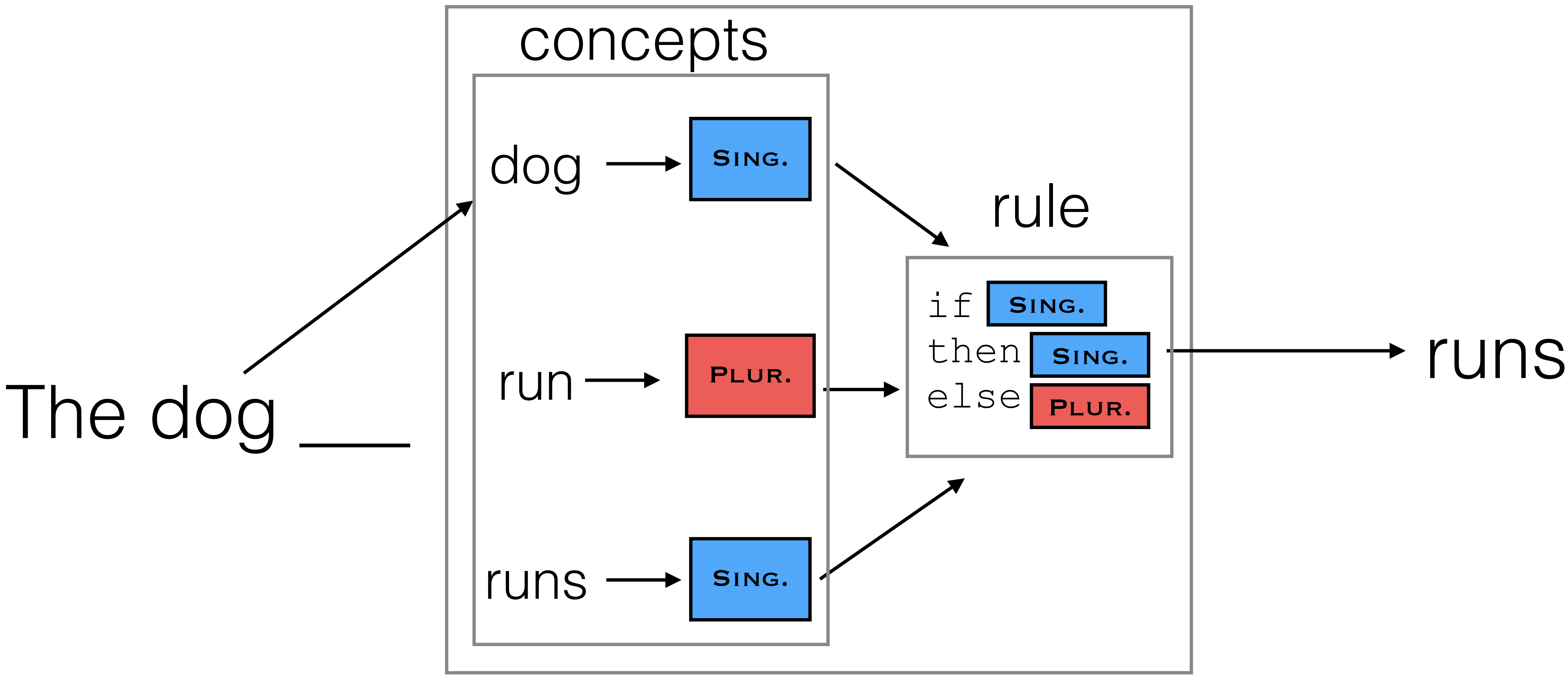
Categories of Reasoning

Idealized Symbolic Learner



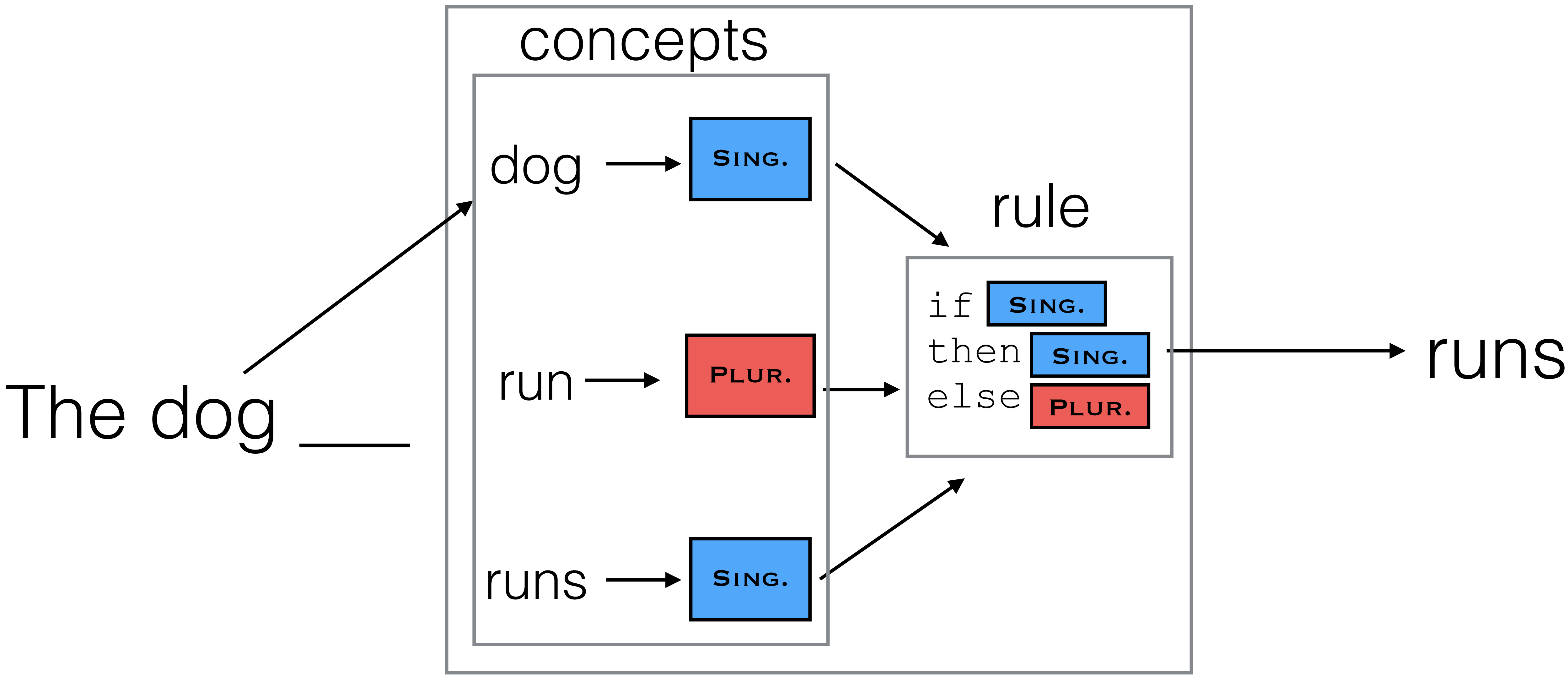
Categories of Reasoning

Idealized Symbolic Learner



Categories of Reasoning

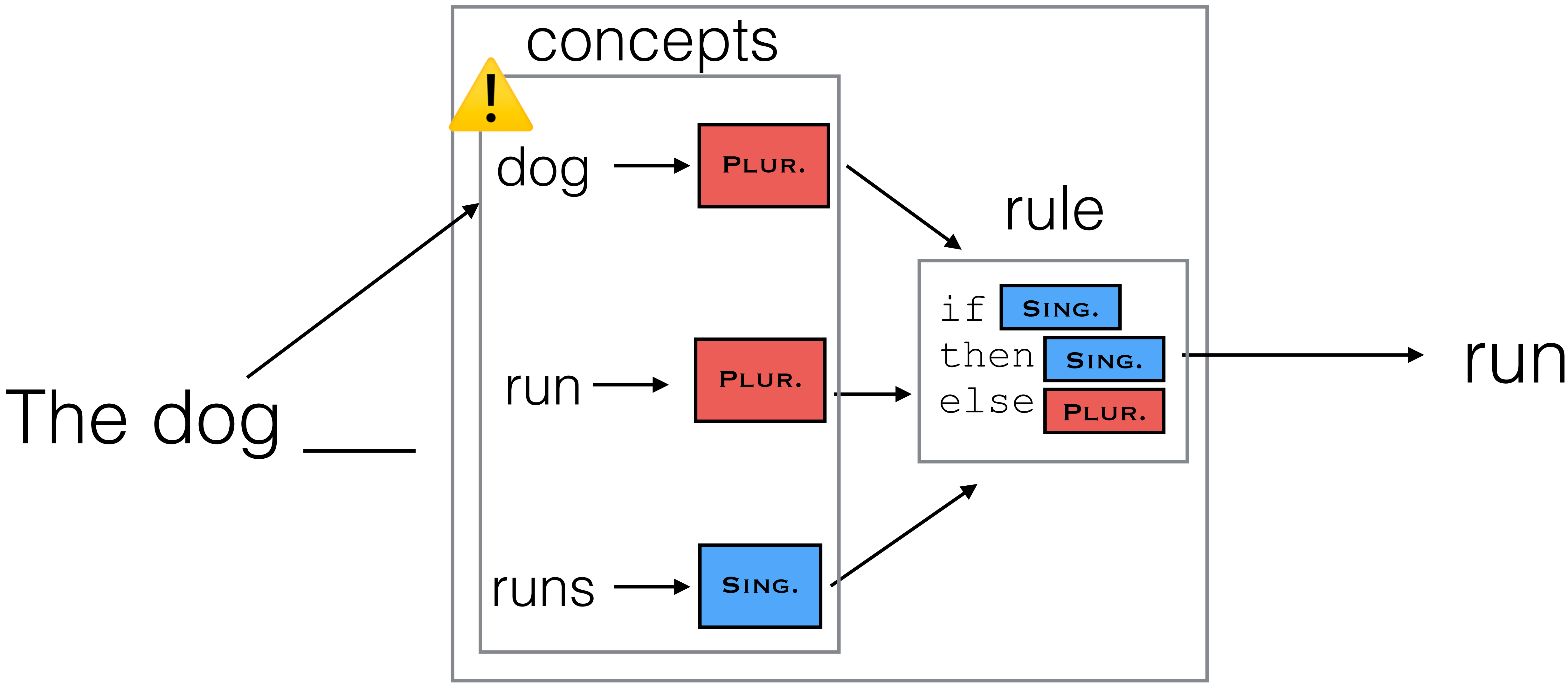
Idealized Symbolic Learner



Decision depends only on abstract concept to which the input is mapped, not on the inputs themselves

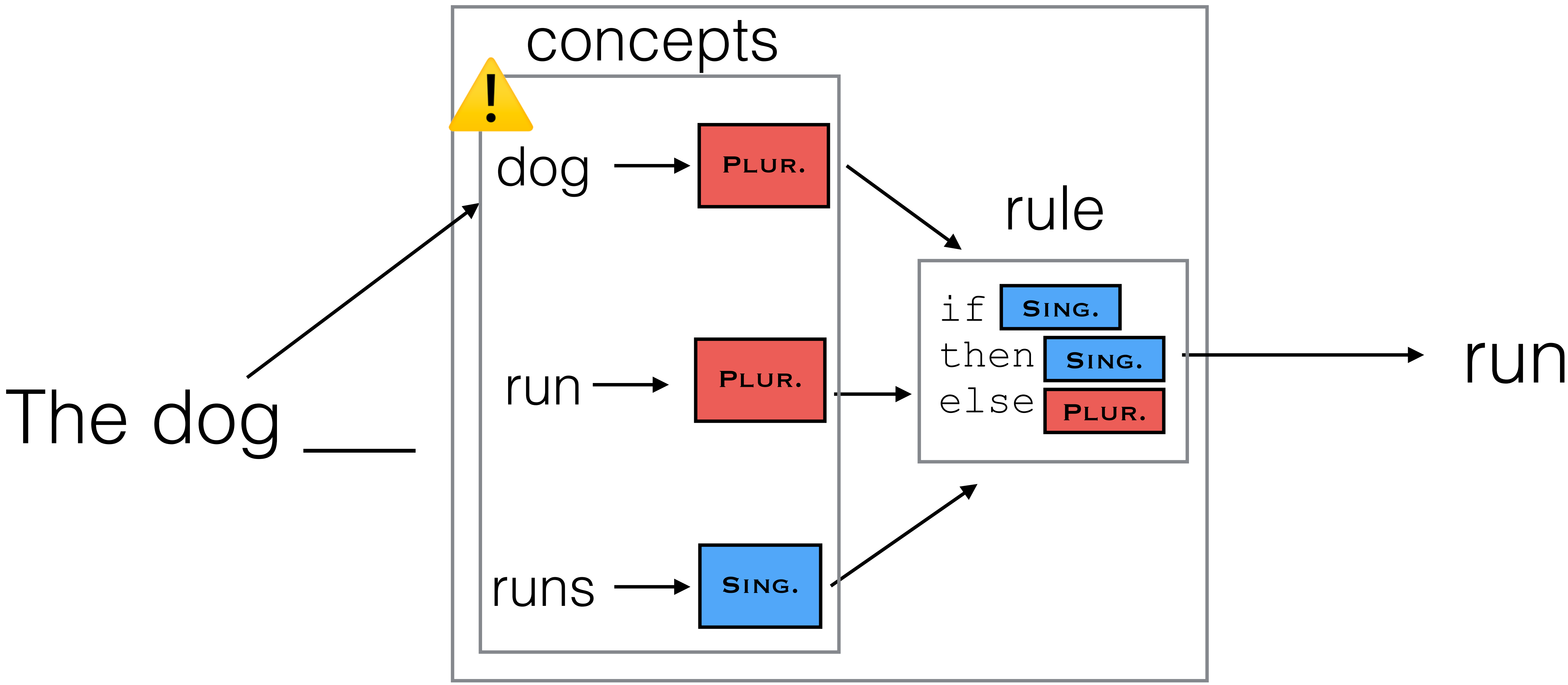
Categories of Reasoning

Symbolic Learner with Noisy Observations



Categories of Reasoning

Symbolic Learner with Noisy Observations



Decision depends only on abstract concept, but mapping from input to concept can be item-specific.

Expectations about Behavior

	Item Specific	Idealized Symbolic	Symbolic + Noisy Obs.

Expectations about Behavior

	Item Specific	Idealized Symbolic	Symbolic + Noisy Obs.
Frequency Effects in Task Performance	✅	⛔	✅

Expectations about Behavior

	Item Specific	Idealized Symbolic	Symbolic + Noisy Obs.
Frequency Effects in Task Performance	✅	⛔	✅
Generalization to Unseen Pairs	⛔	✅	⚠️

Expectations about Behavior

	Item Specific	Idealized Symbolic	Symbolic + Noisy Obs.
Frequency Effects in Task Performance	✅	⛔	✅
Generalization to Unseen Pairs	⛔	✅	⚠️
Task Errors Explained by Observation Errors	⛔	—	✅

Experimental Setup

- Model: BERT trained from scratch on Wikipedia Text (manipulated as needed); no fine-tuning
- IO: The dogs that chase the cat [MASK] fast -> P(run) vs. P(runs)
- Data: Natural and Nonce Sentences:
  - *Addition of such minor **characters** {**seem, seems**} more promotional …*
  - *The **astronomer** of the first session that year {**perform, performs**}…*

Evaluating BERT's Behavior

	Item Specific	Idealized Symbolic	Symbolic + Noisy Obs.	BERT
Frequency Effects in Task Performance	✅	⛔	✅	
Generalization to Unseen Pairs	⛔	✅	⚠️	
Task Errors Explained by Observation Errors	⛔	—	✅	

Frequency Effects in Performance

Effect of Absolute Frequency
(Holding Relative Fixed)

#("runs")

Frequency Effects in Performance

Effect of Absolute Frequency
(Holding Relative Fixed)

$$\#(\text{“runs”})$$

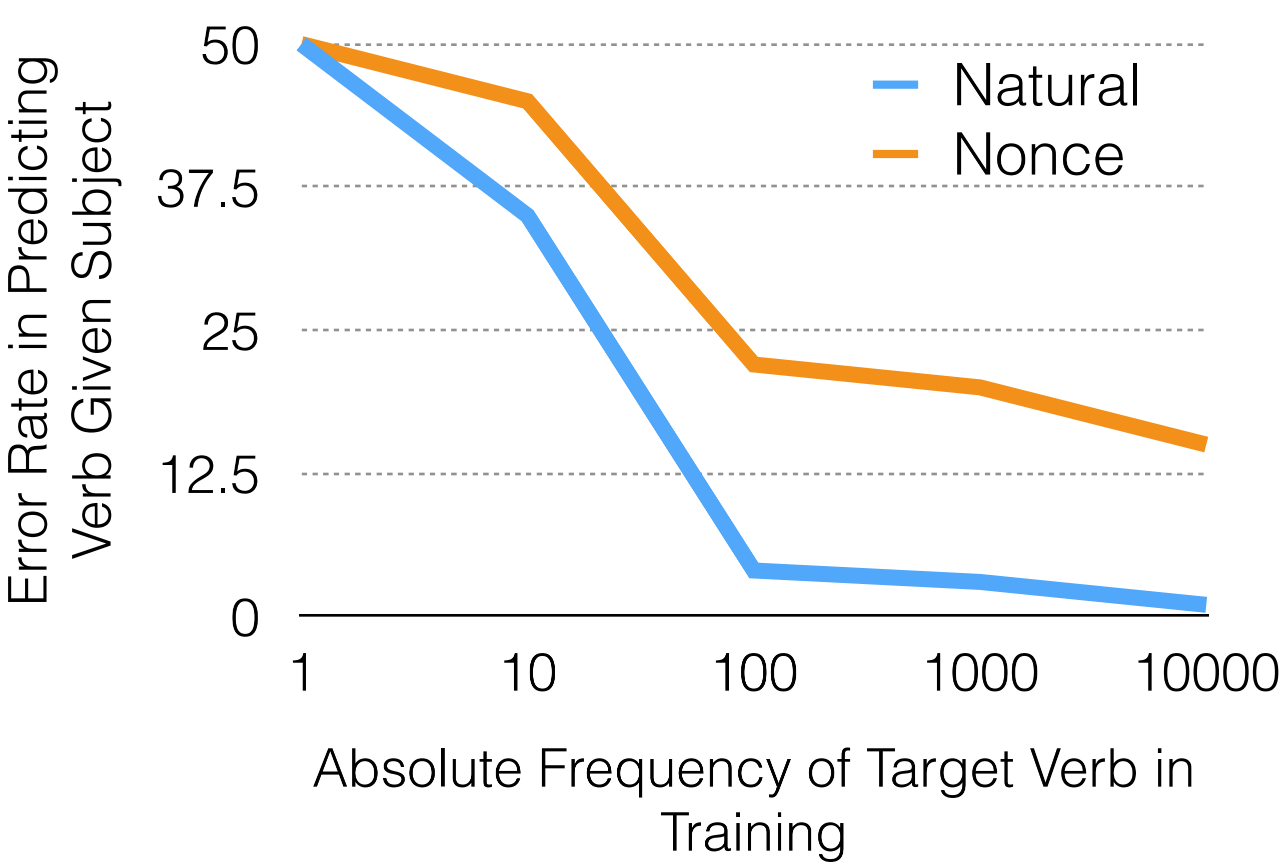
Effect of Relative Frequency
(Holding Absolute Fixed)

$$\frac{\#(\text{“runs”})}{\#(\text{“run”})}$$

Frequency Effects in Performance

Effect of Absolute Frequency
(Holding Relative Fixed)

#("runs")



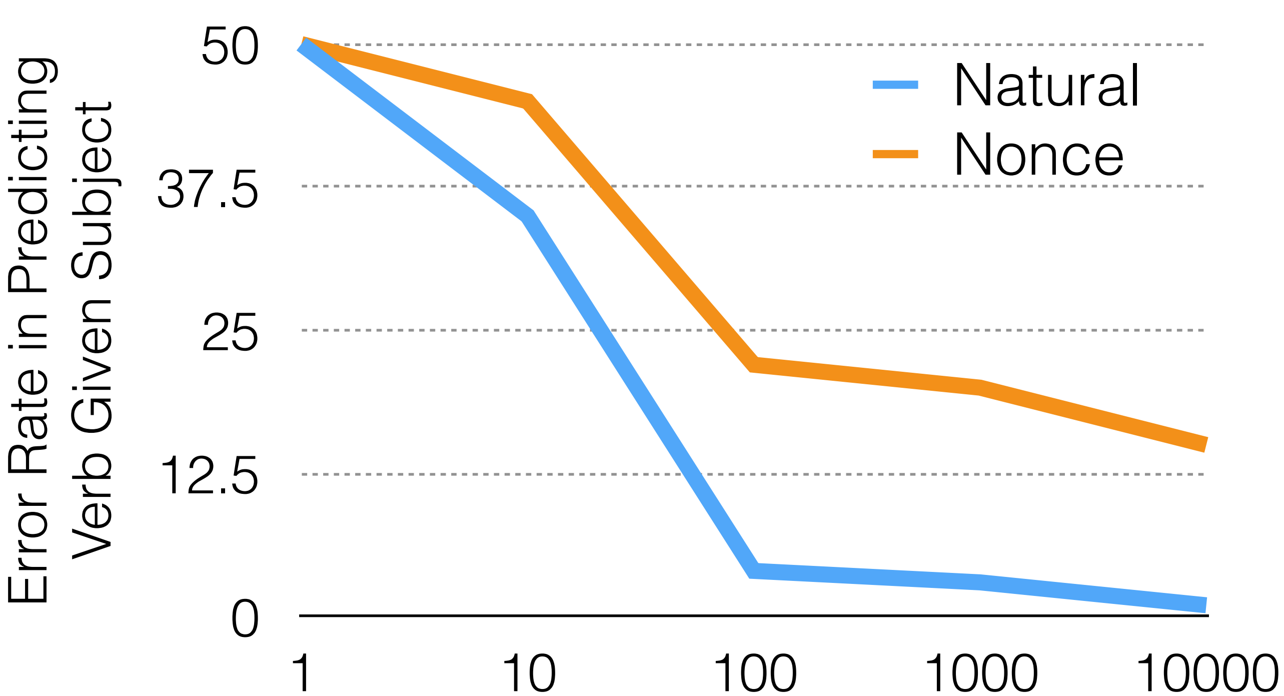
Effect of Relative Frequency
(Holding Absolute Fixed)

$$\frac{\#(\text{"runs"})}{\#(\text{"run"})}$$

Frequency Effects in Performance

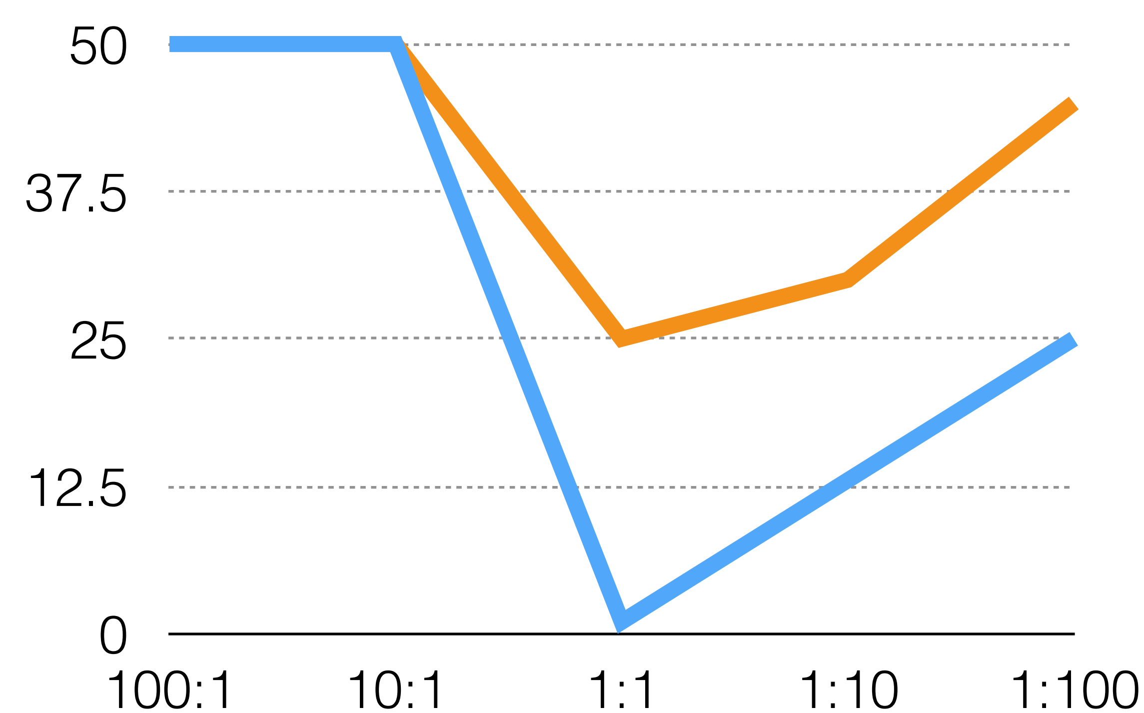
Effect of Absolute Frequency
(Holding Relative Fixed)

#("runs")



Effect of Relative Frequency
(Holding Absolute Fixed)

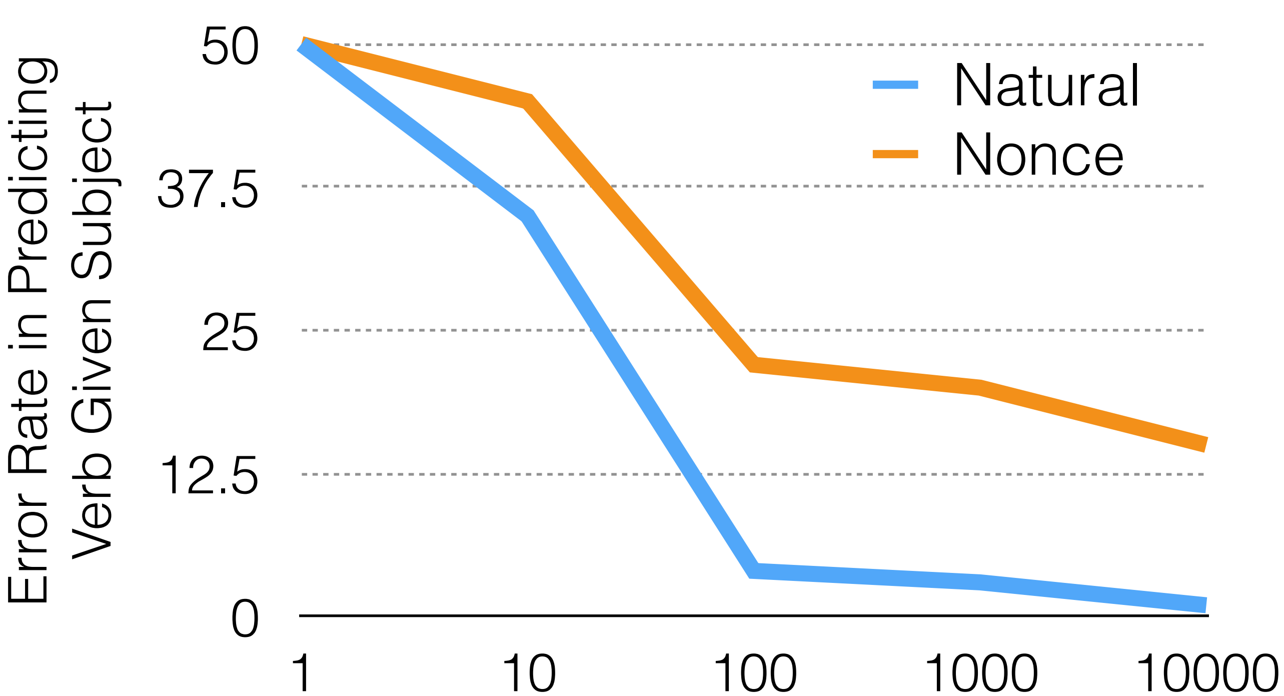
$$\frac{\#(\text{"runs"})}{\#(\text{"run"})}$$



Frequency Effects in Performance

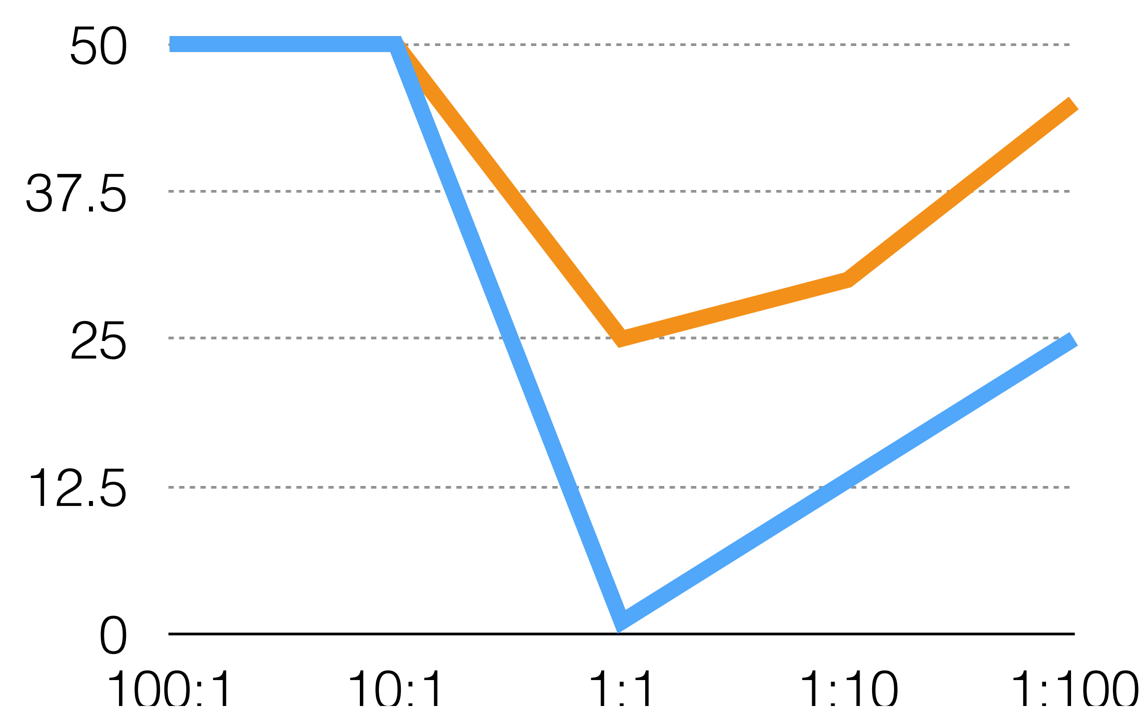
Effect of Absolute Frequency
(Holding Relative Fixed)

#("runs")



Effect of Relative Frequency
(Holding Absolute Fixed)

$$\frac{\#(\text{"runs"})}{\#(\text{"run"})}$$

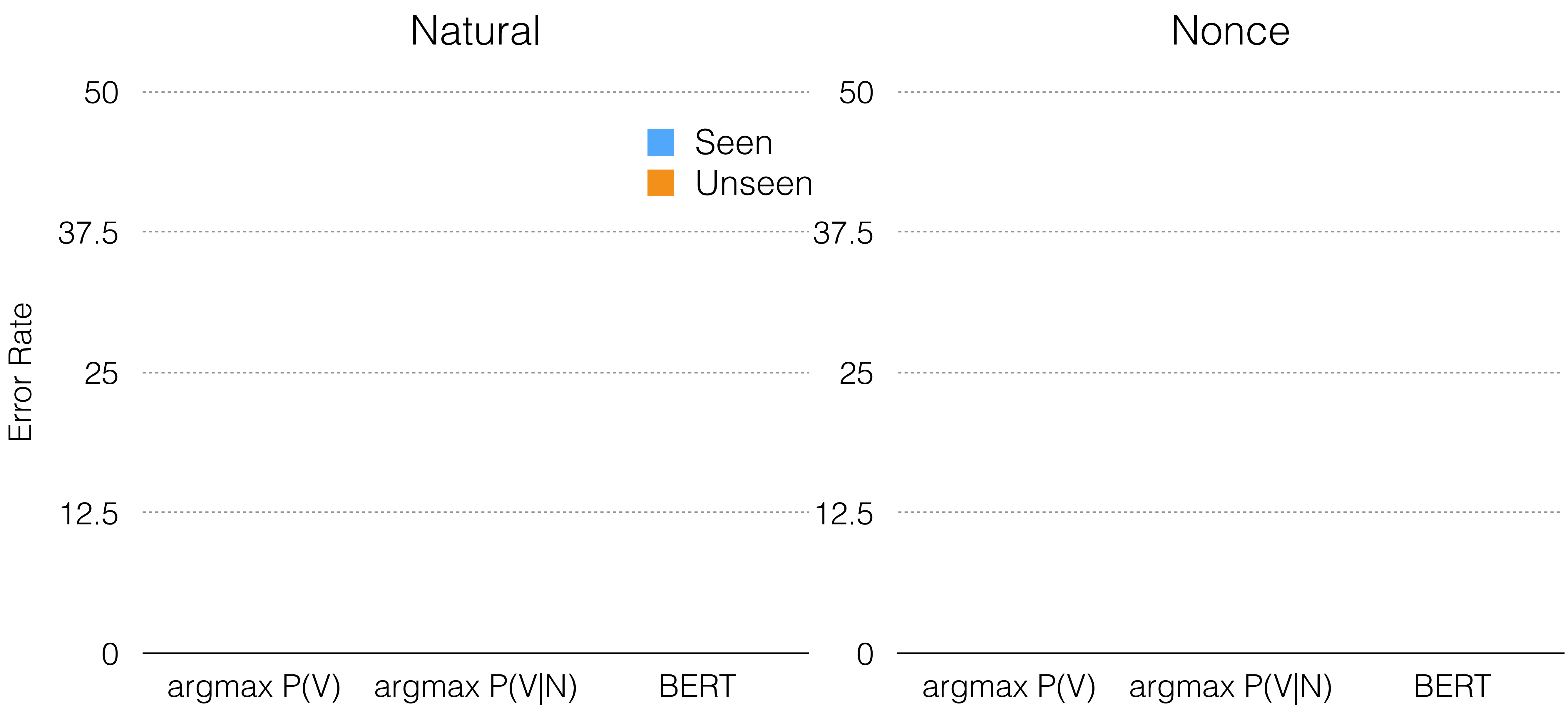


Both the absolute and the relative frequency of an item independently influence model performance.

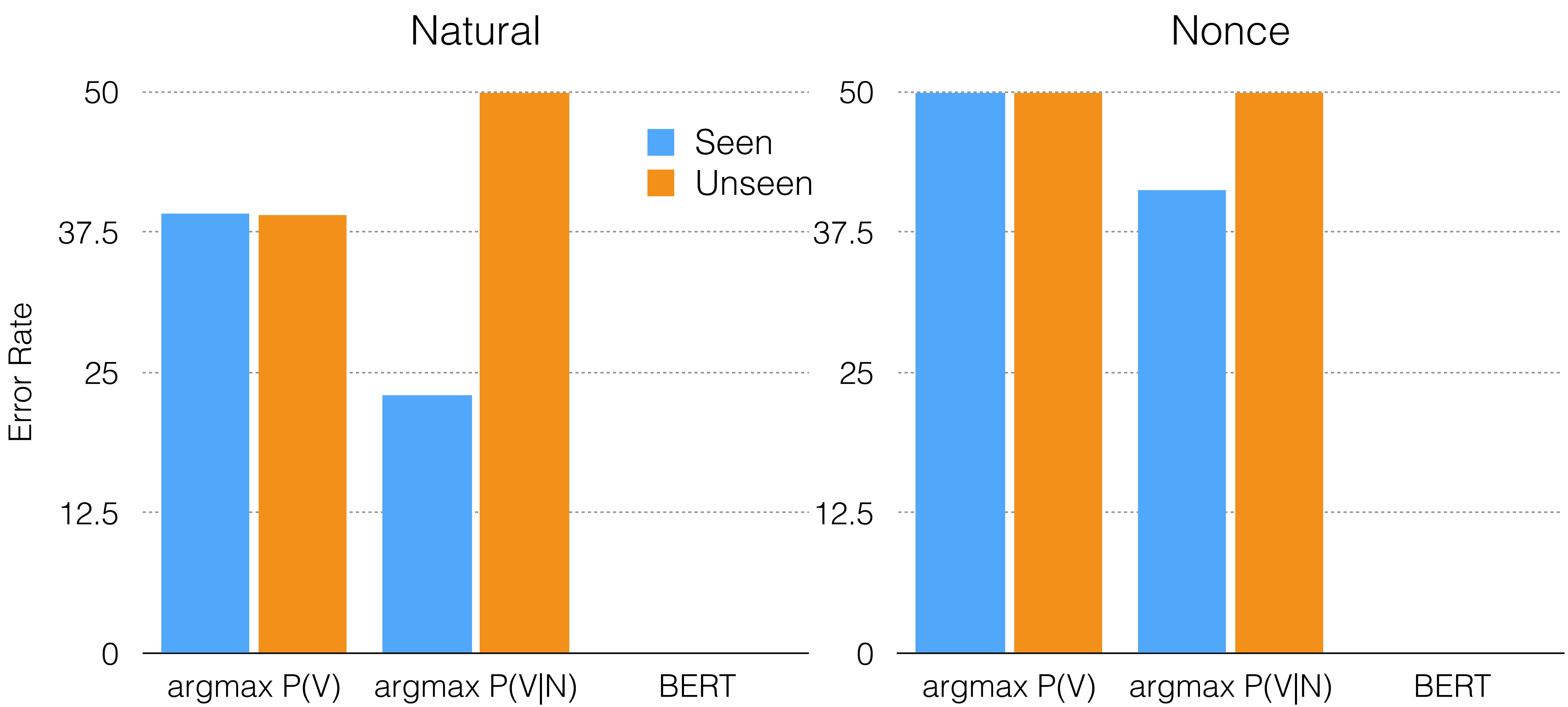
Evaluating BERT's Behavior

	Item Specific	Idealized Symbolic	Symbolic + Noisy Obs.	BERT
Frequency Effects in Task Performance	✅	⛔	✅	✅
Generalization to Unseen Pairs	⛔	✅	⚠️	
Task Errors Explained by Observation Errors	⛔	—	✅	

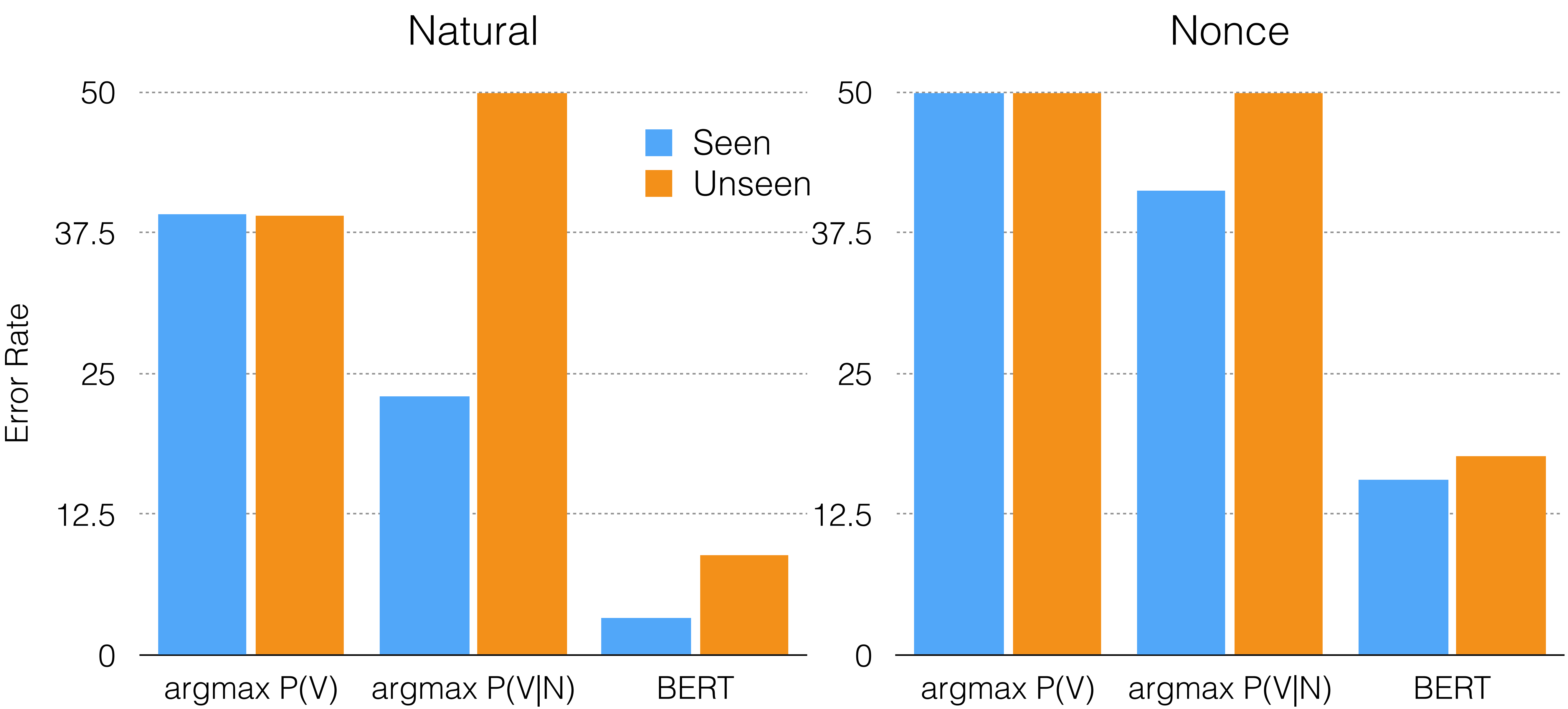
Generalization to Unseen Noun-Verb Pairs



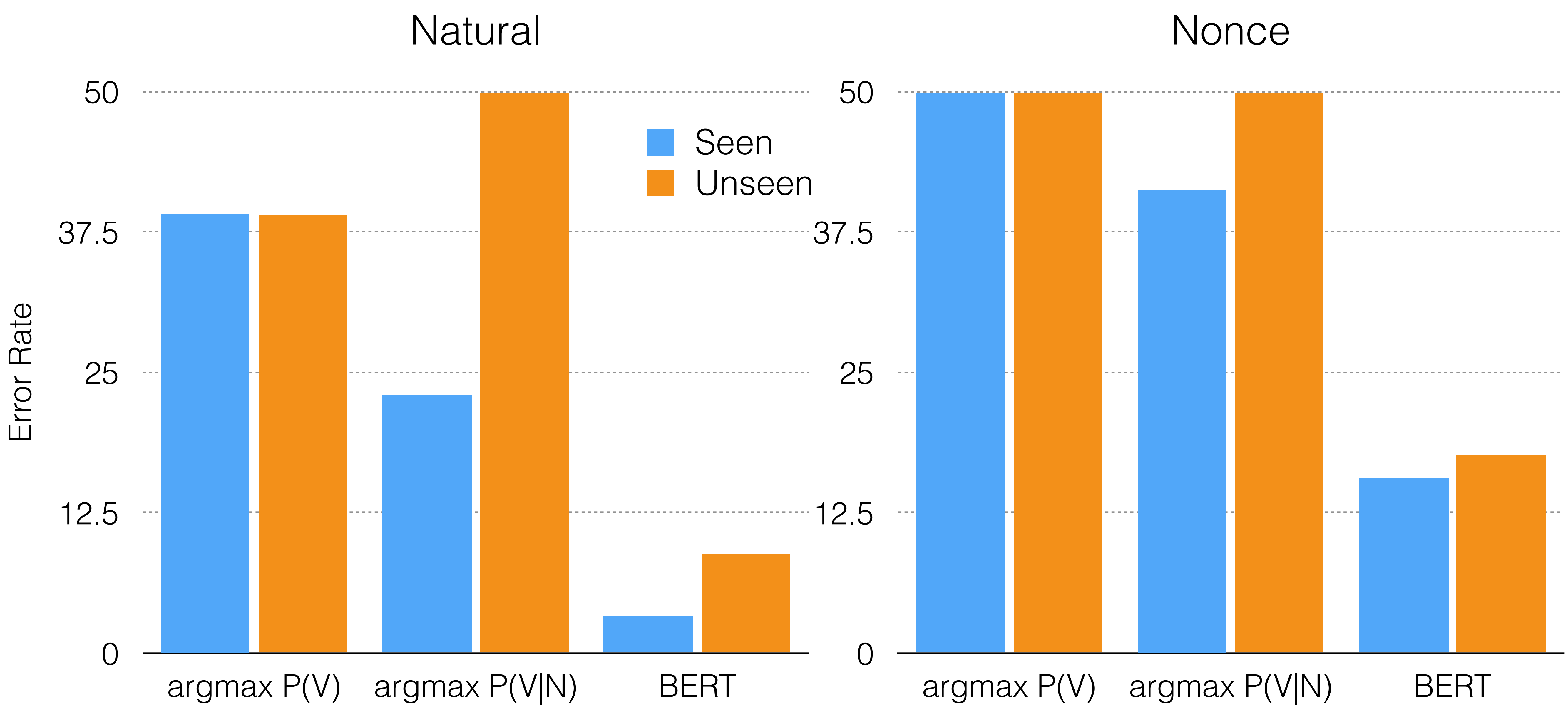
Generalization to Unseen Noun-Verb Pairs



Generalization to Unseen Noun-Verb Pairs



Generalization to Unseen Noun-Verb Pairs



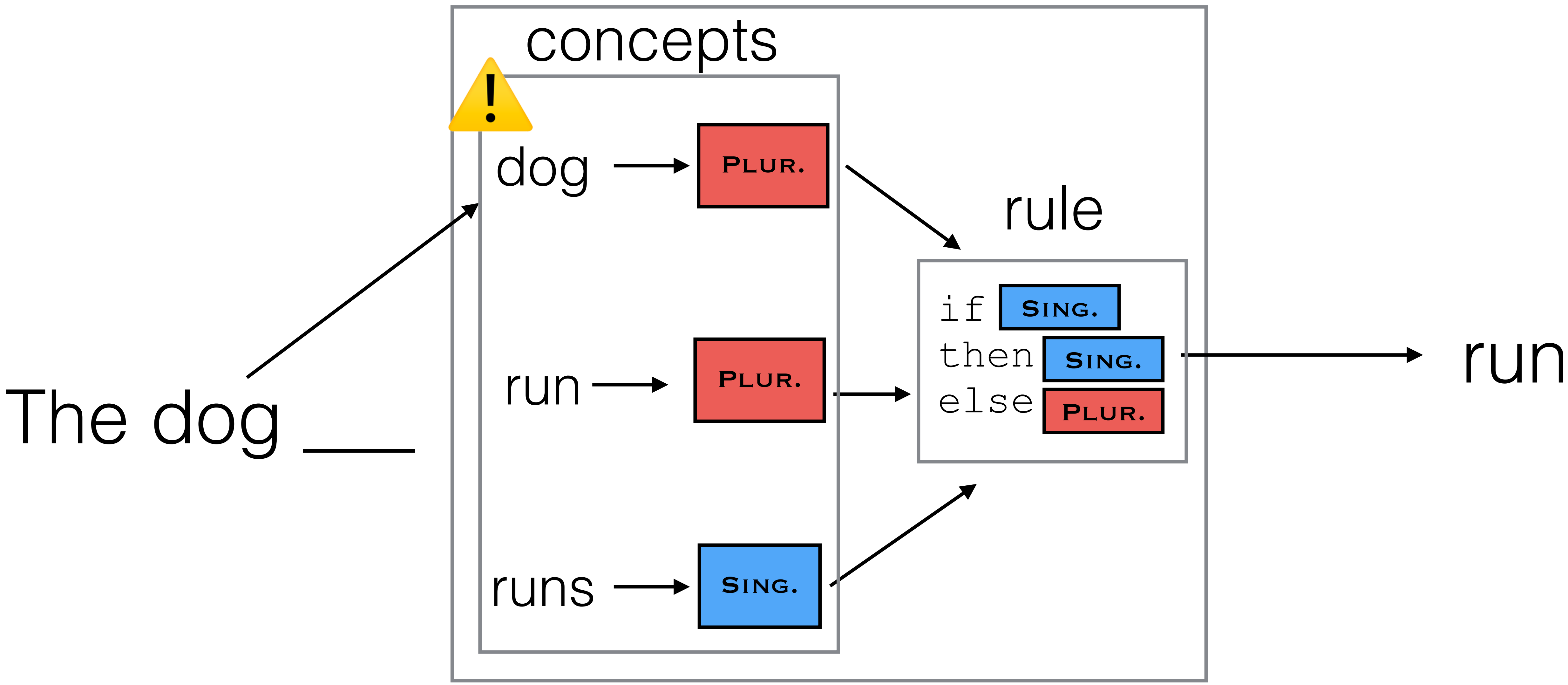
Small increase in error on unseen items, but well below error expect for purely item-specific procedure.

Evaluating BERT's Behavior

	Item Specific	Idealized Symbolic	Symbolic + Noisy Obs.	BERT
Frequency Effects in Task Performance	✅	⛔	✅	✅
Generalization to Unseen Pairs	⛔	✅	⚠️	⚠️
Task Errors Explained by Observation Errors	⛔	—	✅	

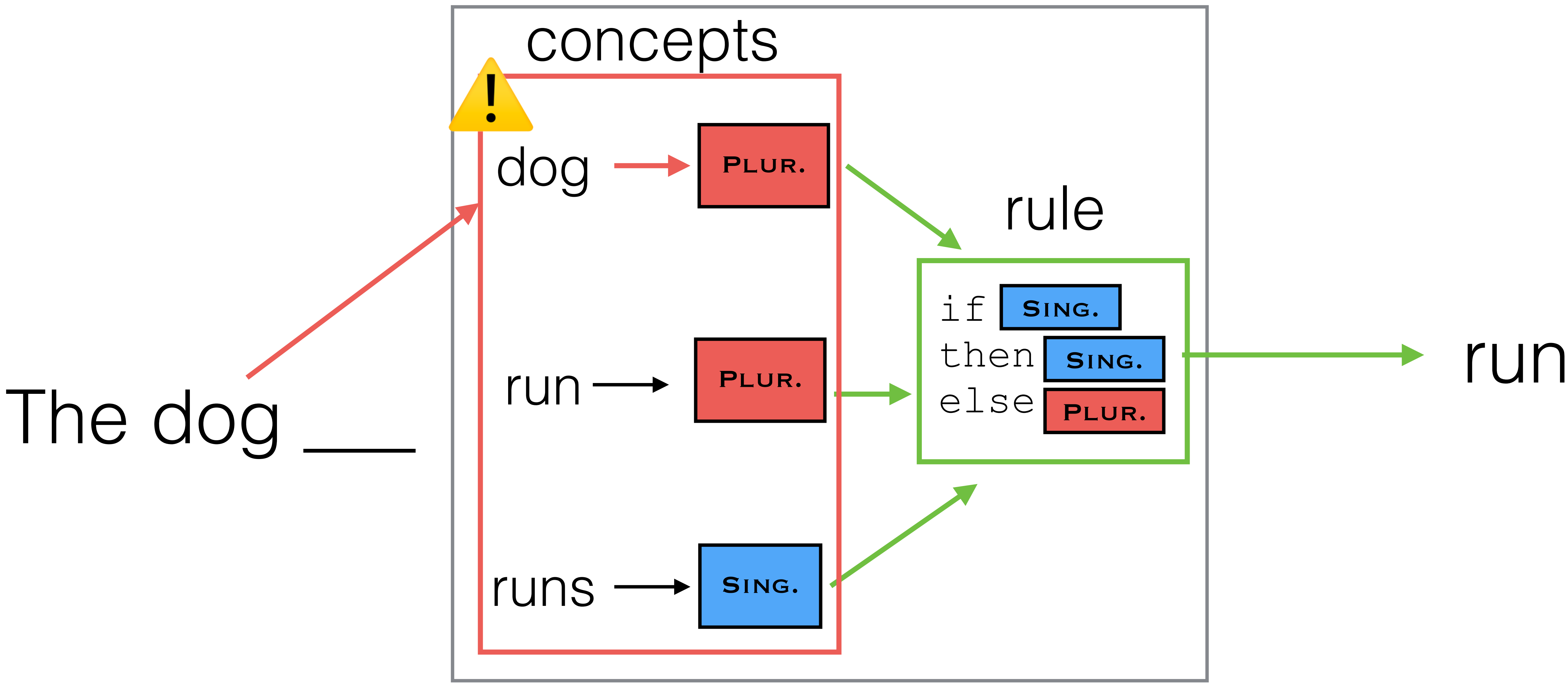
Categories of Reasoning

Symbolic Learner with Noisy Observations



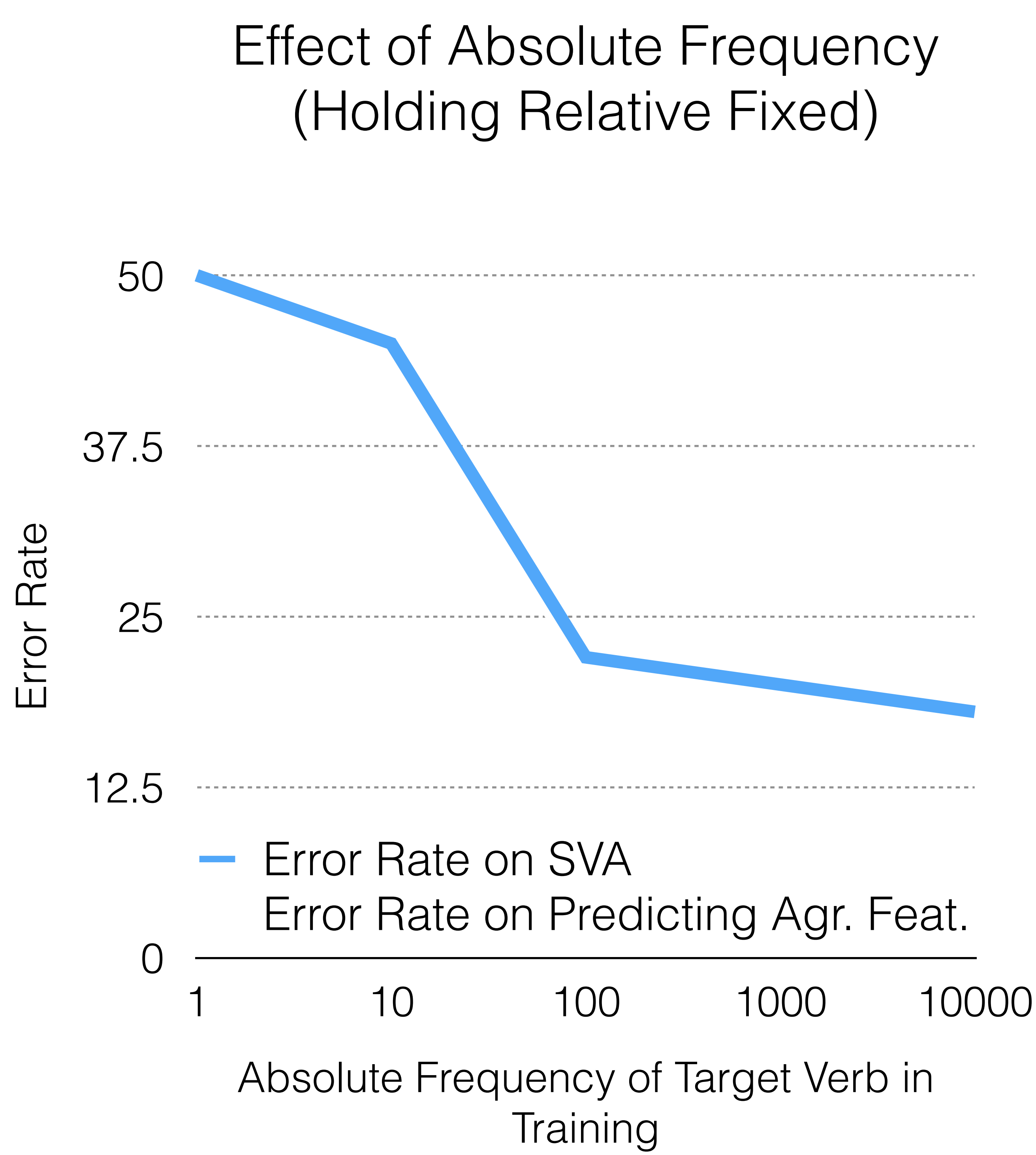
Categories of Reasoning

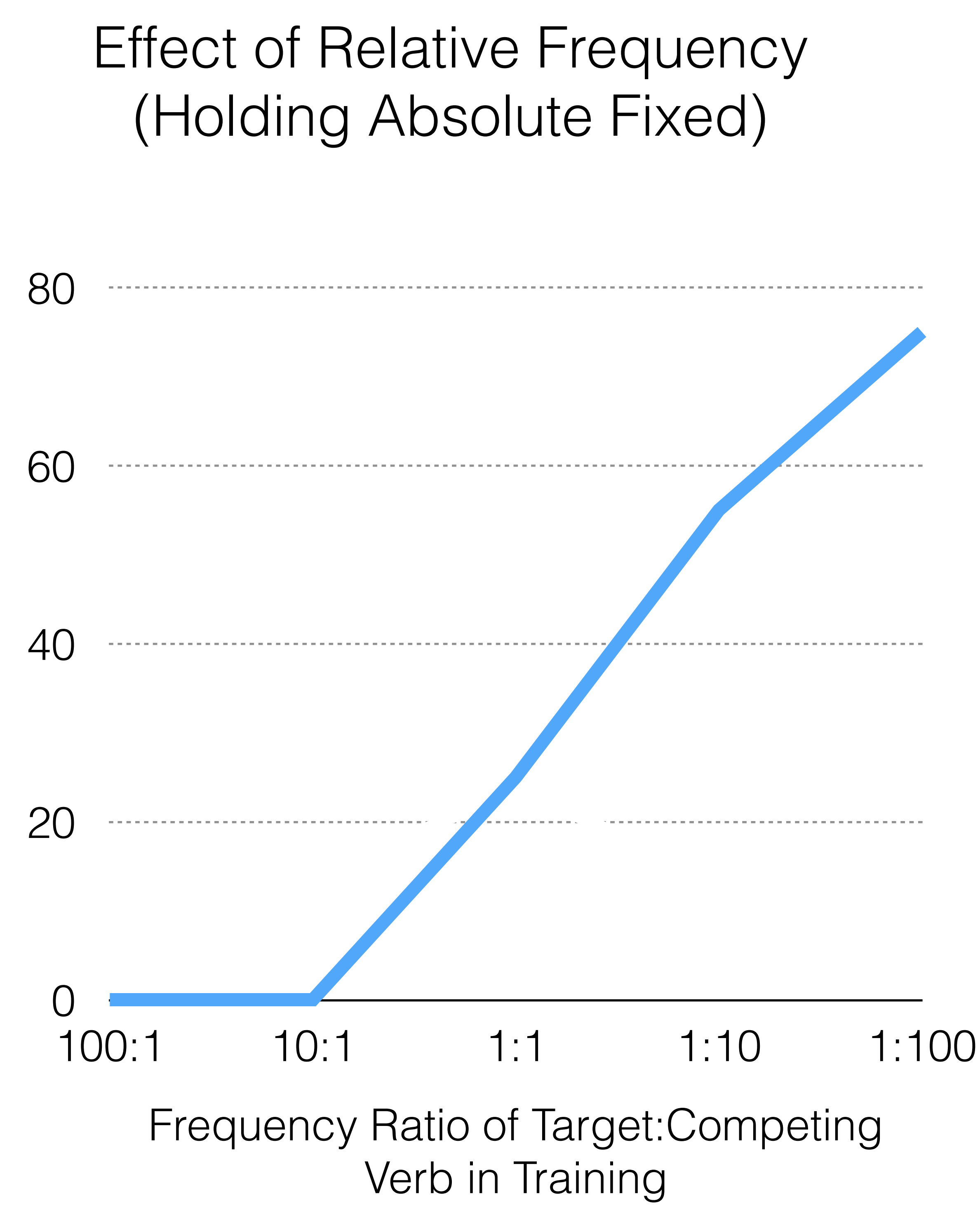
Symbolic Learner with Noisy Observations



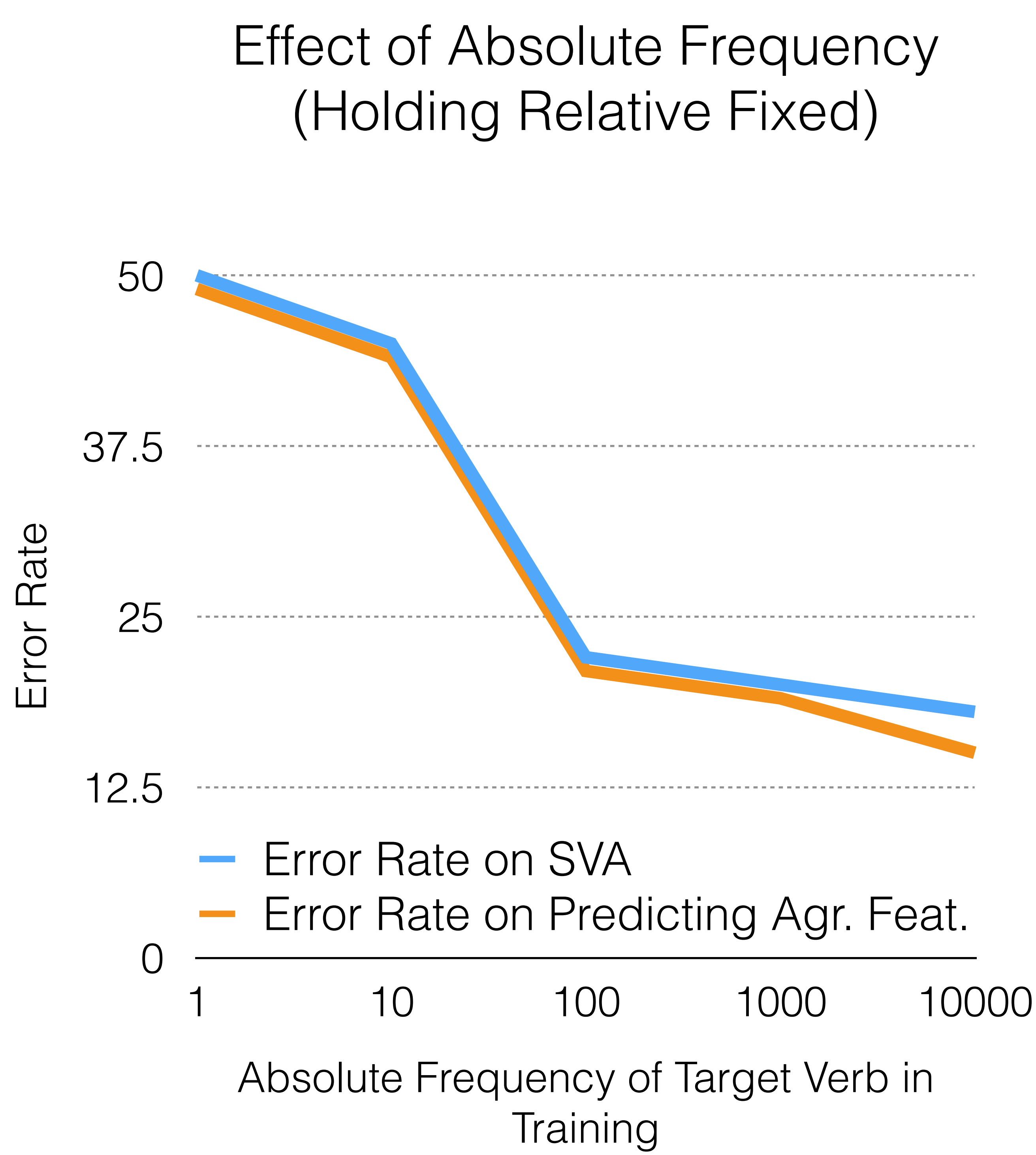
Errors in output are due to errors in mapping inputs to concepts, not errors in rule.

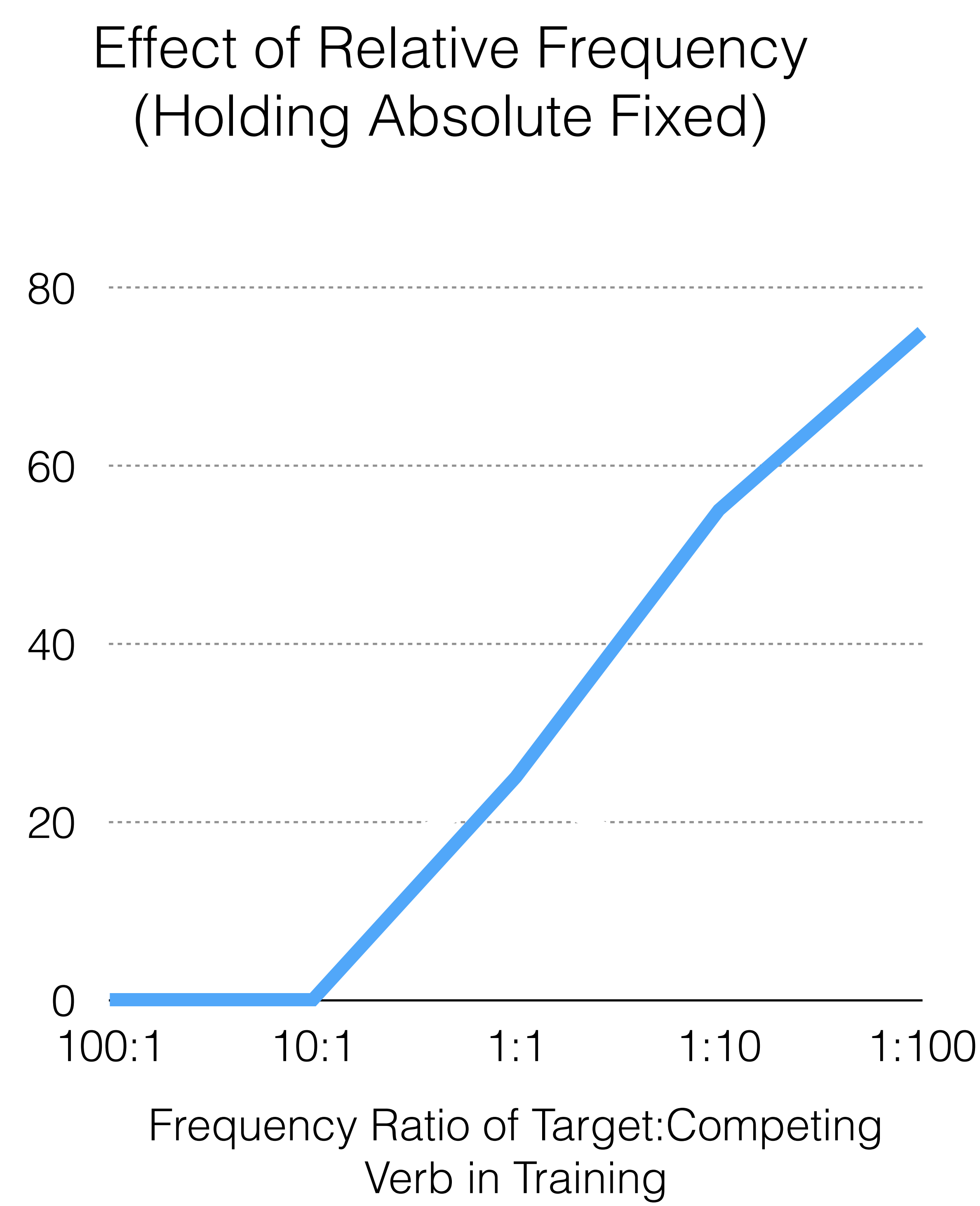
Frequency Effects Explained by Errors in Agreement Feature?



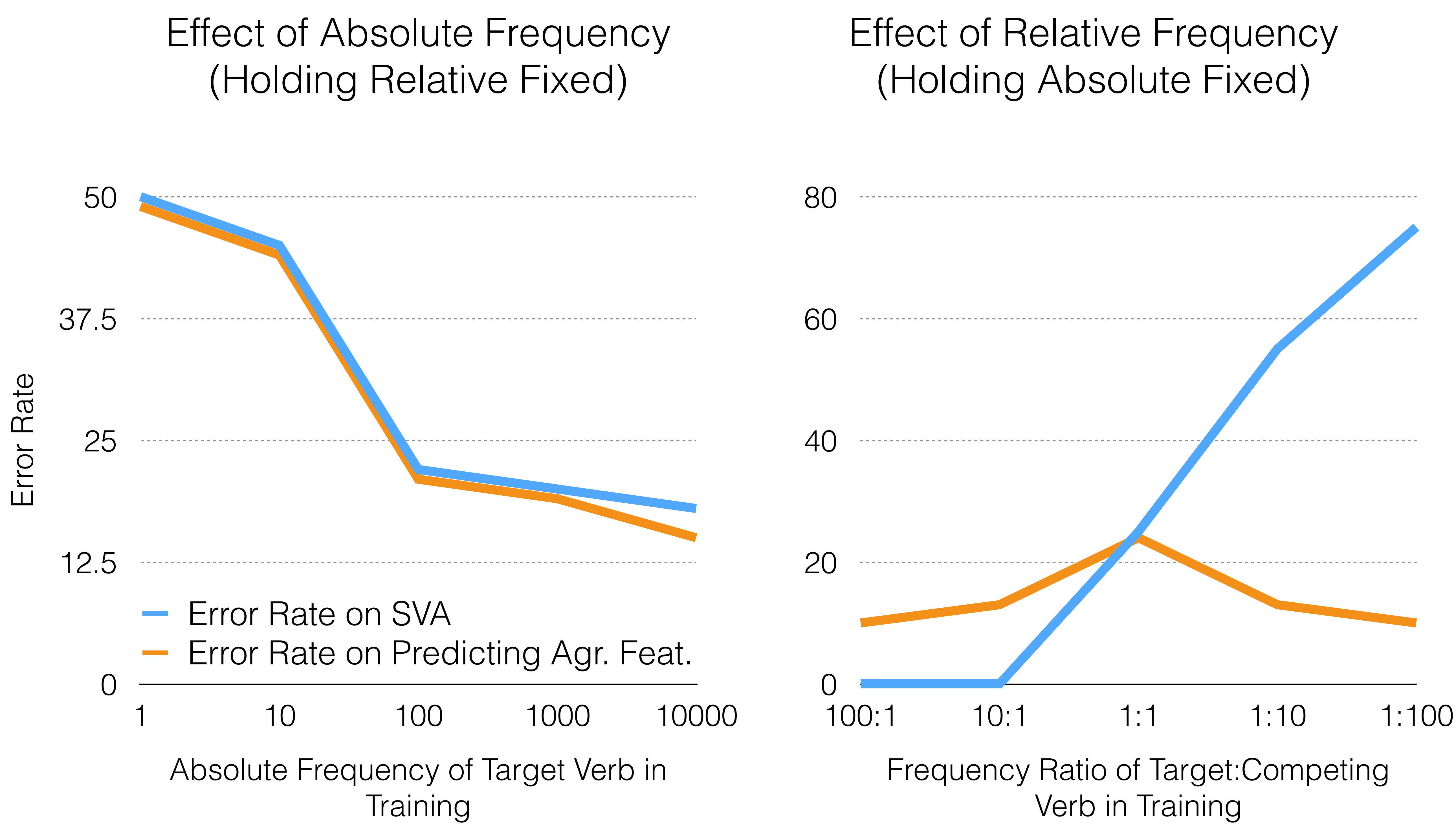


Frequency Effects Explained by Errors in Agreement Feature?





Frequency Effects Explained by Errors in Agreement Feature?



Evaluating BERT's Behavior

	Item Specific	Idealized Symbolic	Symbolic + Noisy Obs.	BERT
Frequency Effects in Task Performance	✅	⛔	✅	✅
Generalization to Unseen Pairs	⛔	✅	⚠️	⚠️
Task Errors Explained by Observation Errors	⛔	—	✅	⚠️

Takeaways

Takeaways

- Pretrained Neural LMs (BERT) exhibit a **mix** of systematic generalization and item-specific memorization

Takeaways

- Pretrained Neural LMs (BERT) exhibit a **mix** of systematic generalization and item-specific memorization
- Behavior is consistent with a model is capable applying **rules to abstract categories**

Takeaways

- Pretrained Neural LMs (BERT) exhibit a **mix** of systematic generalization and item-specific memorization
- Behavior is consistent with a model is capable applying **rules to abstract categories**
- **BUT** still behaves incorrectly when:

Takeaways

- Pretrained Neural LMs (BERT) exhibit a **mix** of systematic generalization and item-specific memorization

- Behavior is consistent with a model is capable applying **rules to abstract categories**

- **BUT** still behaves incorrectly when:

  - It **incorrectly classifies inputs** according to the concepts. Appears related to abs. frequency.

Takeaways

- Pretrained Neural LMs (BERT) exhibit a **mix** of systematic generalization and item-specific memorization
- Behavior is consistent with a model is capable applying **rules to abstract categories**
- **BUT** still behaves incorrectly when:
  - It **incorrectly classifies inputs** according to the concepts. Appears related to abs. frequency.
  - It has to **overcome strong priors**. Related to rel. frequency.

Takeaways

- Pretrained Neural LMs (BERT) exhibit a **mix** of systematic generalization and item-specific memorization

- Behavior is consistent with a model is capable applying **rules to abstract categories**

- **BUT** still behaves incorrectly when:

    - It **incorrectly classifies inputs** according to the concepts. Appears related to abs. frequency.

    - It has to **overcome strong priors**. Related to rel. frequency. (See Lovering et al, 2021)

Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. [TACL 2022]

Jason Wei, Dan Garrette, Tal Linzen and Ellie Pavlick. Frequency Effects on Syntactic Rule Learning in Transformers. [EMNLP 2021]

Charles Lovering, **Rohan Jha**, Tal Linzen and Ellie Pavlick. Predicting Inductive Biases of Pretrained Models. [ICLR 2021]

Aaron Traylor, Roman Feiman and Ellie Pavlick. AND does not mean OR: Using Formal Languages to Study Language Models' Representations. [ACL 2021]

Roma Patel and Ellie Pavlick. Mapping Language Models to Grounded Conceptual Spaces. [ICLR 2022]

General Discussion

General Discussion

- We should think of “symbolic reasoning” as a **computation-level phenomenon**

General Discussion

- We should think of "symbolic reasoning" as a **computation-level phenomenon**
- It is in-principle possible for neural networks to be **functionally equivalent** to the models we traditionally think of as "symbolic reasoners" in cognitive and computer—e.g., BayesNets

General Discussion

- We should think of "symbolic reasoning" as a **computation-level phenomenon**

- It is in-principle possible for neural networks to be **functionally equivalent** to the models we traditionally think of as "symbolic reasoners" in cognitive and computer—e.g., BayesNets

- Diagnosing whether this is the case for modern NNs requires multifaceted evaluations that focus on **representations, not just behavior**

General Discussion

- We should think of “symbolic reasoning” as a **computation-level phenomenon**

- It is in-principle possible for neural networks to be **functionally equivalent** to the models we traditionally think of as “symbolic reasoners” in cognitive and computer—e.g., BayesNets

- Diagnosing whether this is the case for modern NNs requires multifaceted evaluations that focus on **representations, not just behavior**

- Progress requires interdisciplinary collaboration and **hypothesis-driven research** on why NNs produce the outputs they do for a given input

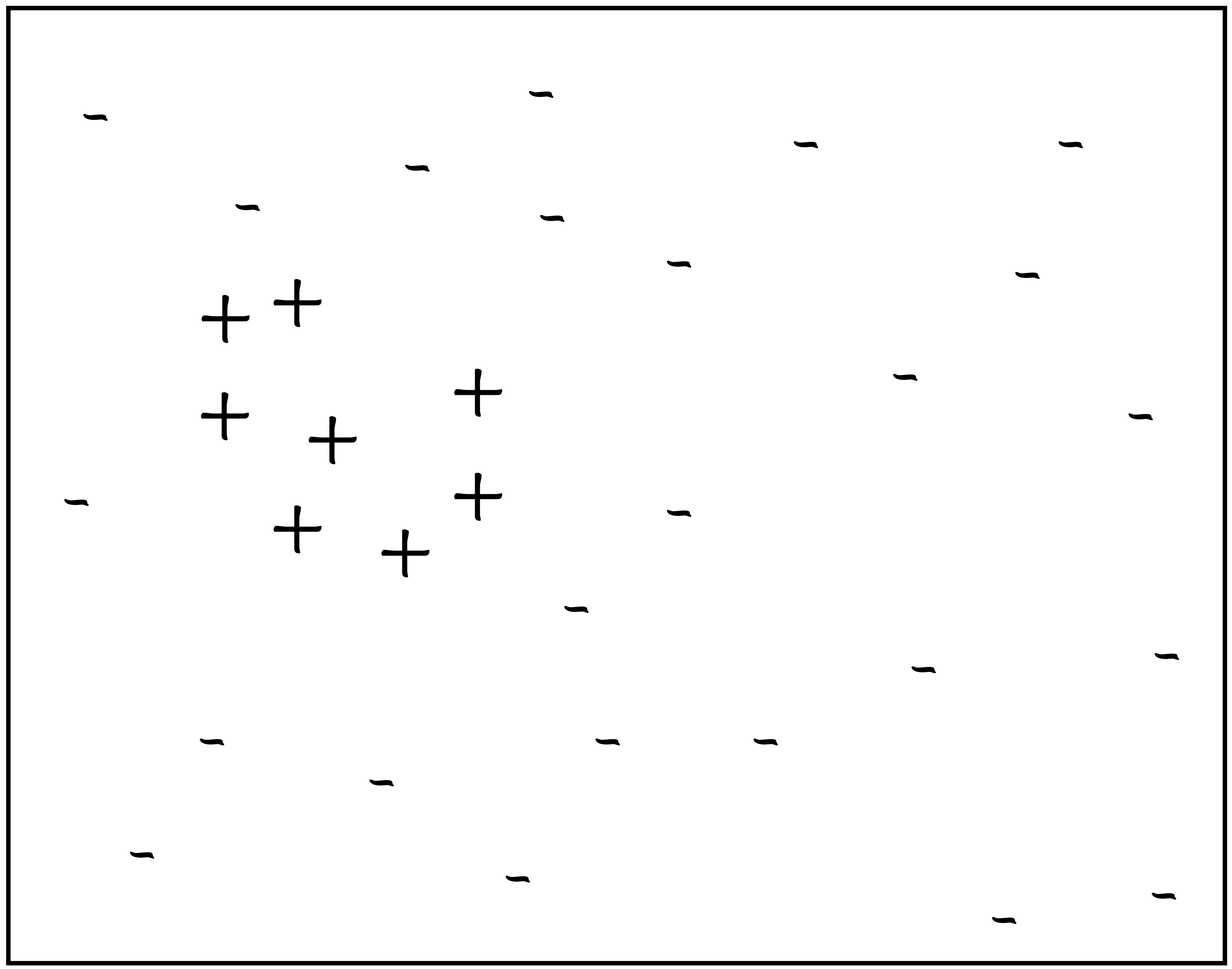
Thank you!

Backup Slides

Question

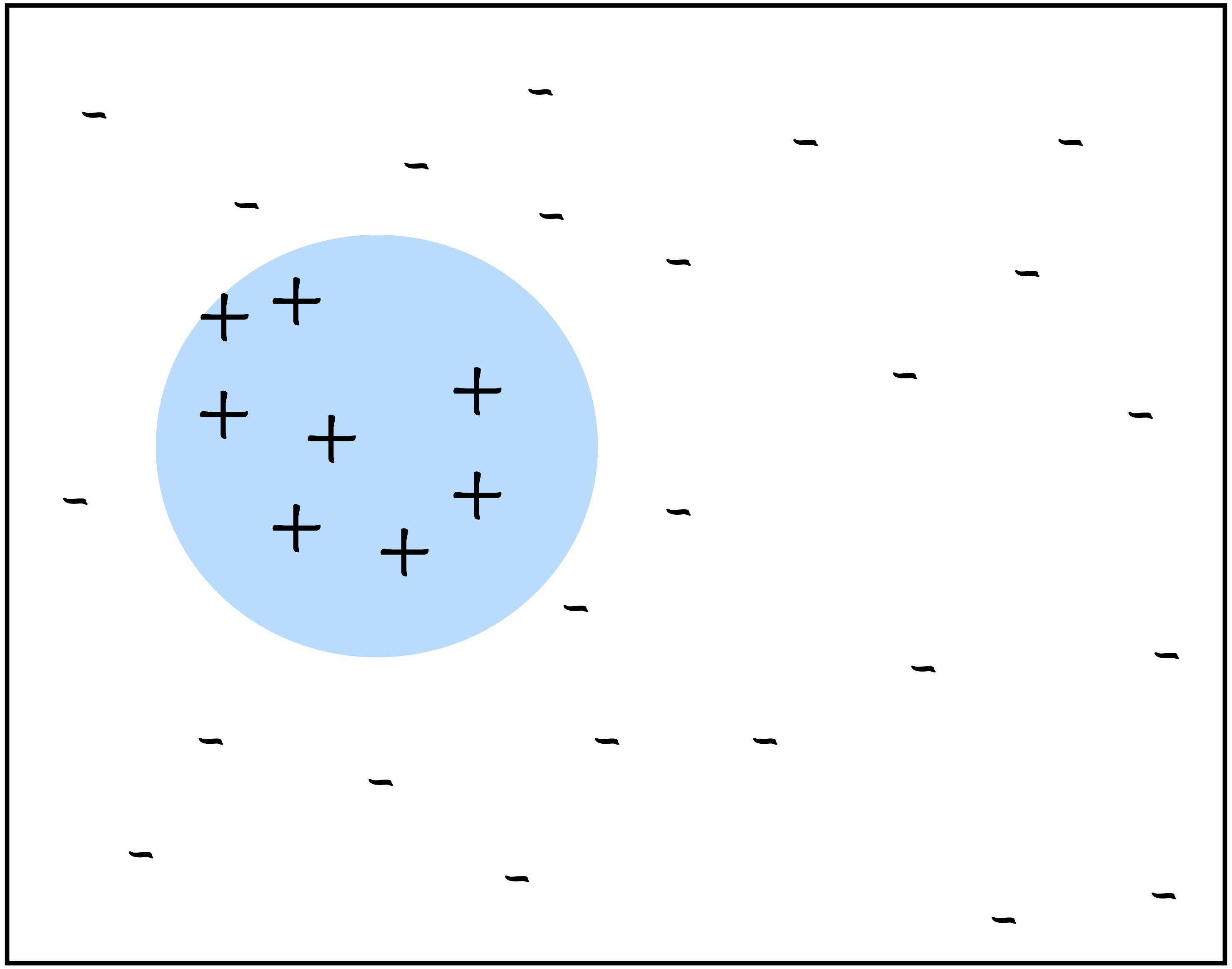
- Can we predict whether or not a given concept will influence a model's predictions based on:
  - The training data?
  - The model's representations?
  - Some combination of the above?

General Set Up

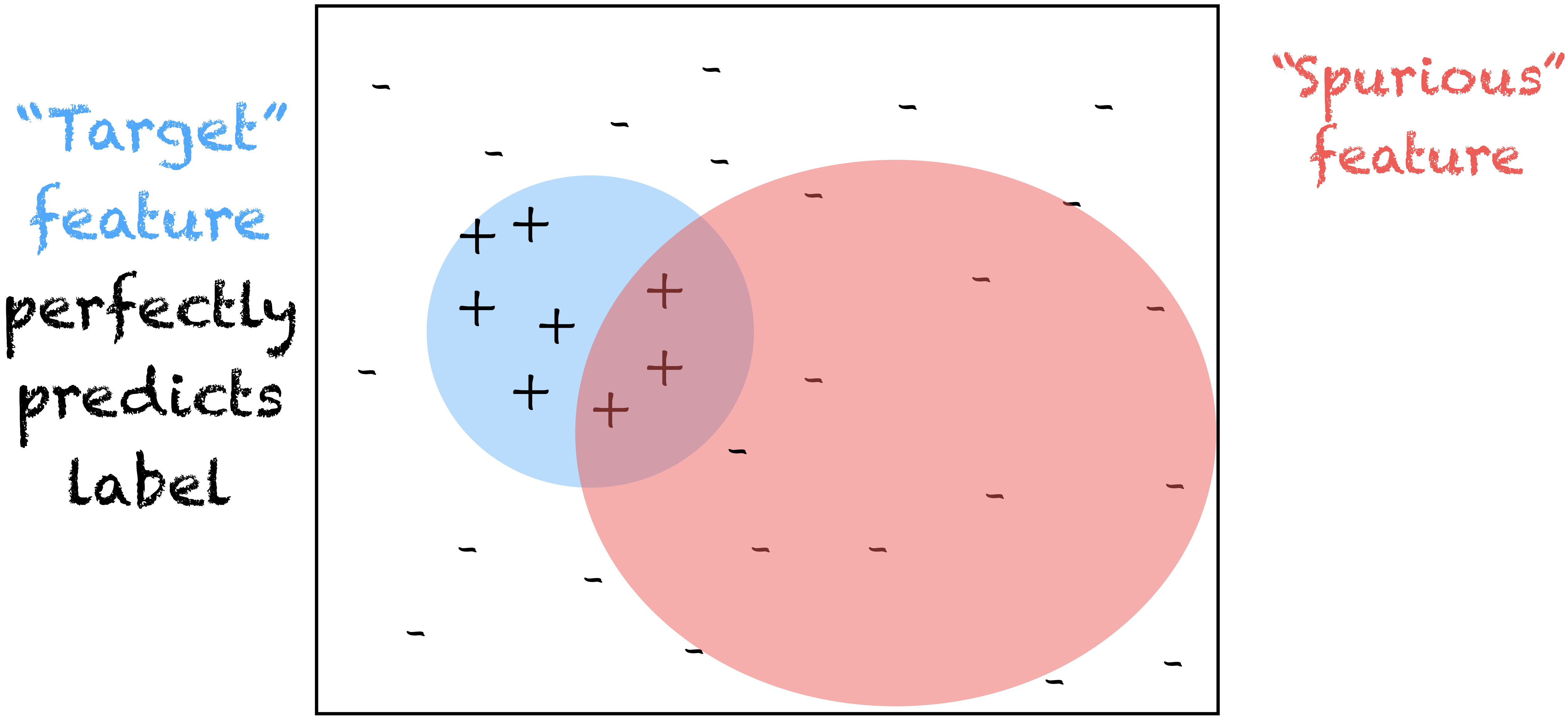


General Set Up

"Target" feature perfectly predicts label

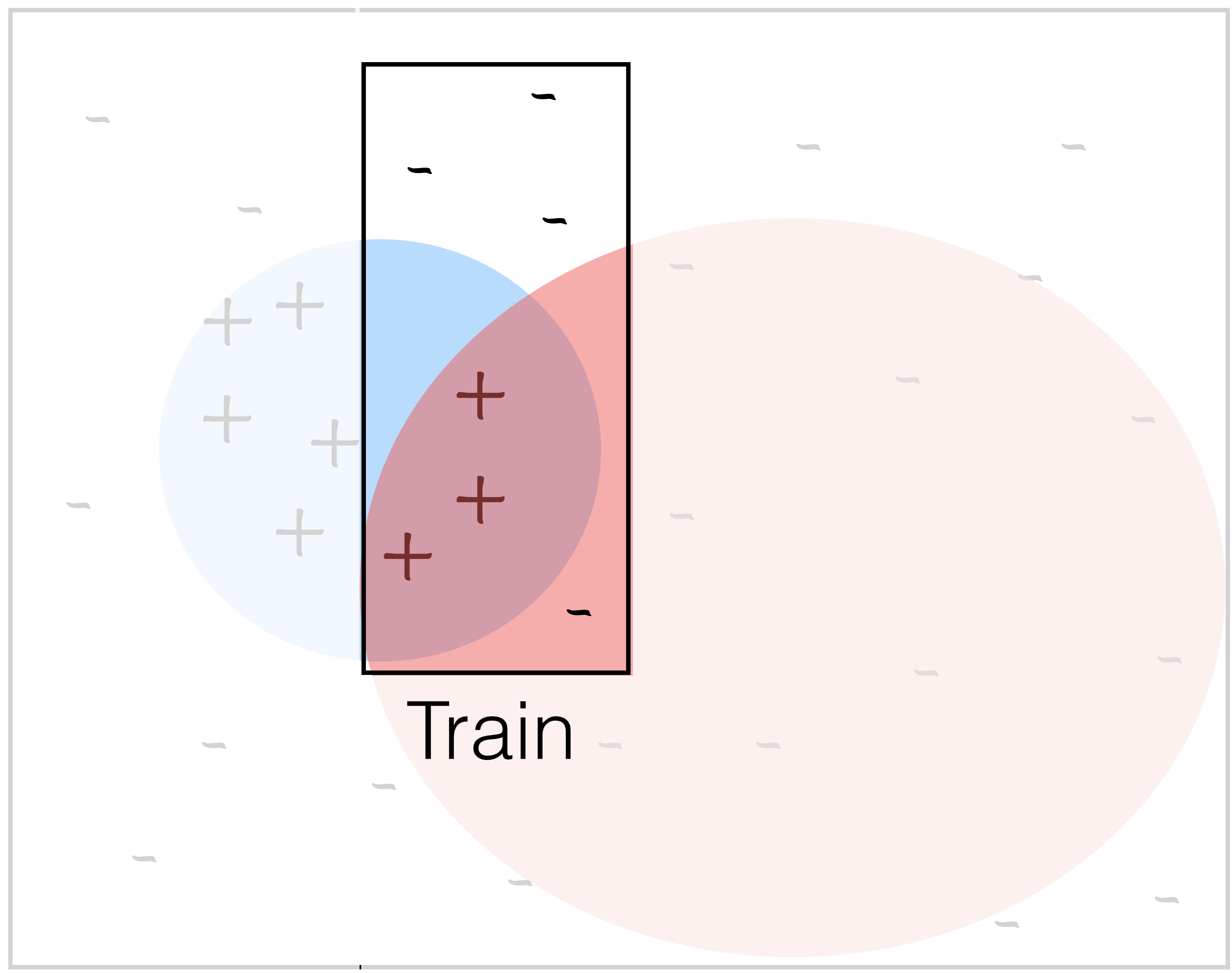


General Set Up



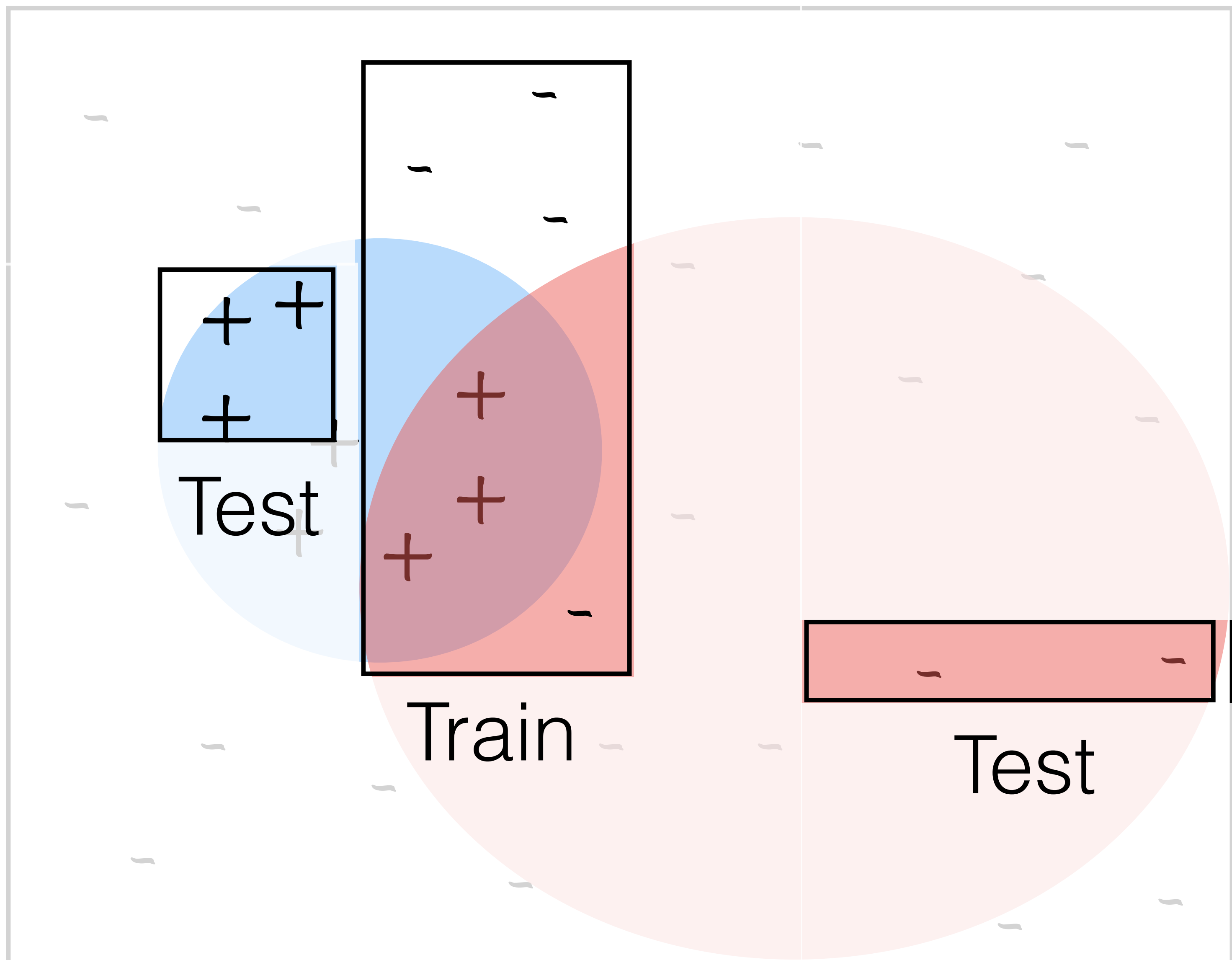
General Set Up

"Target" feature perfectly predicts label



"Spurious" feature which happens to co-occur with target in training sample

General Set Up



Generalizing well out of training distribution requires using the target feature

Toy Sentence Classification Task

Name	Target	Spurious	Example
`contains-1`	a '1' occurs in the sequence	a '2' occurs in the sequence	2 4 11 1 4
`prefix-duplicate`	sequence begins with a duplicate	a '2' occurs in the sequence	5 5 11 12 2
`adjacent-duplicate`	duplicate occurs somewhere in the sequence	a '2' occurs in the sequence	11 12 3 3 2
`first-last`	first symbol and last symbol are the same	a '2' occurs in the sequence	7 2 11 12 7

Predicting Inductive Biases of Pretrained Models
Jha, Lovering, Linzen, and Pavlick (2020)

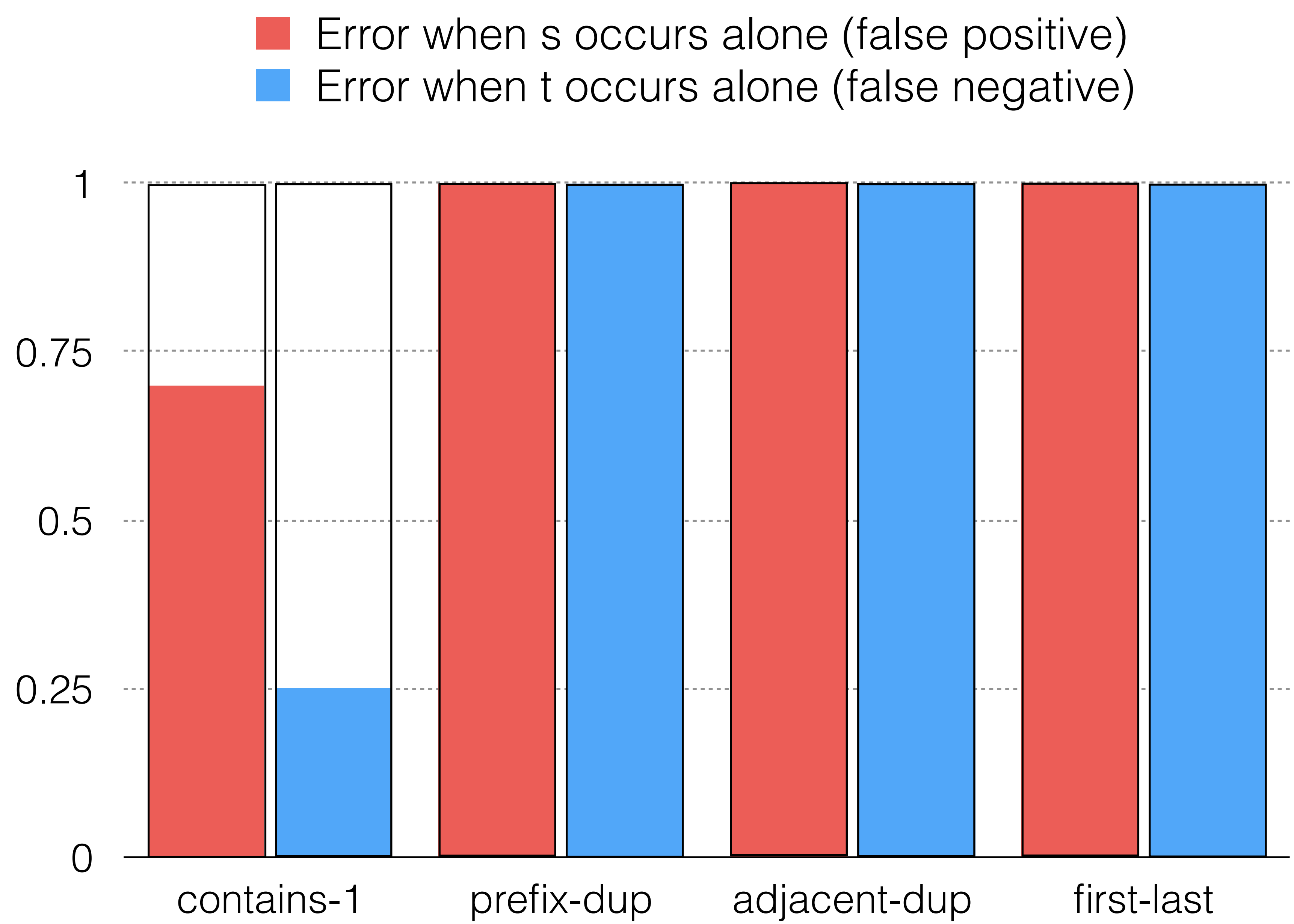
Out-of-Distribution Test Error

Perfect co-occurrence between spurious and target

Predicting Inductive Biases of Pretrained Models
Jha, Lovering, Linzen, and Pavlick (2020)

Out-of-Distribution Test Error

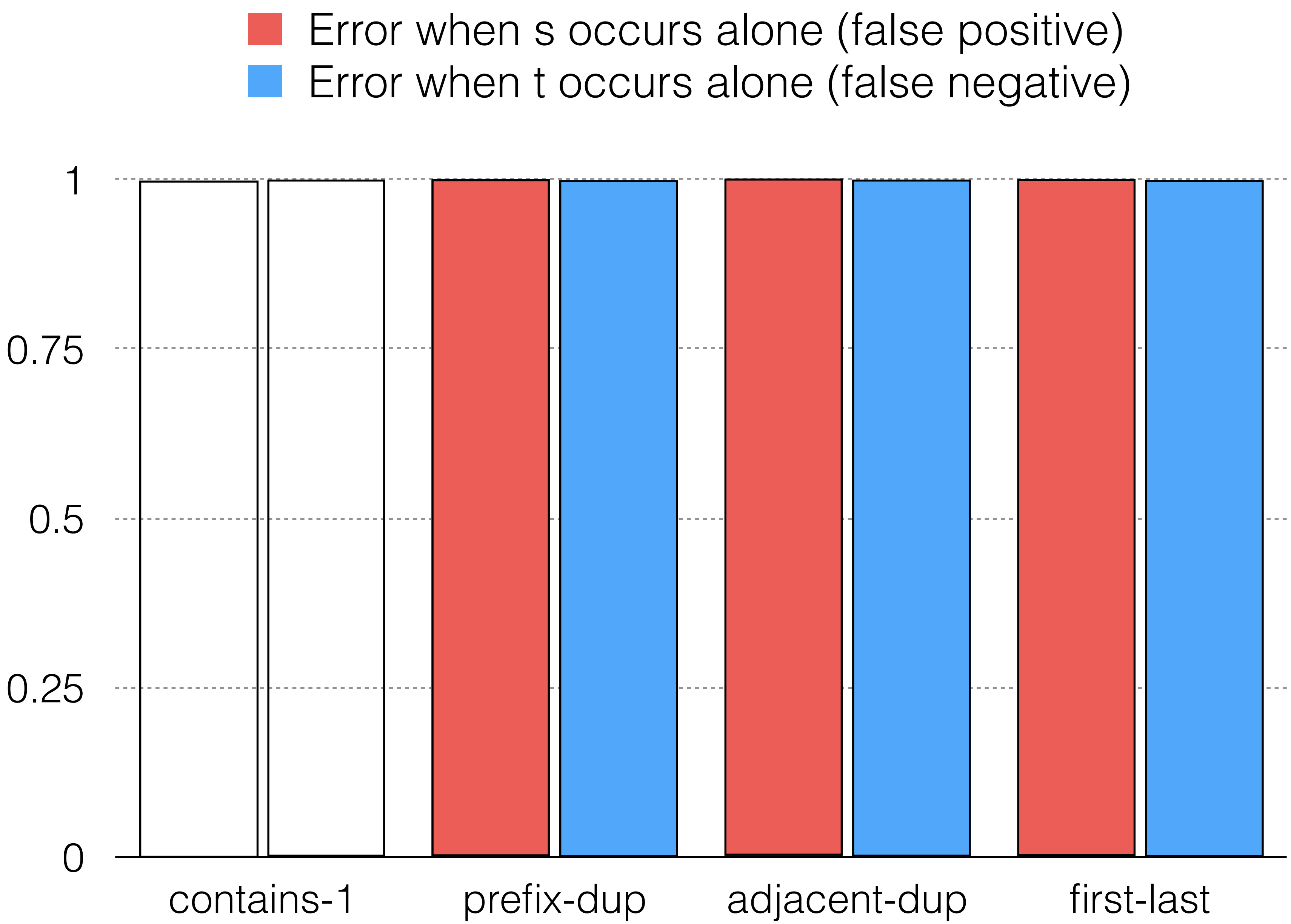
Perfect co-occurrence between spurious and target



Predicting Inductive Biases of Pretrained Models
Jha, Lovering, Linzen, and Pavlick (2020)

Out-of-Distribution Test Error

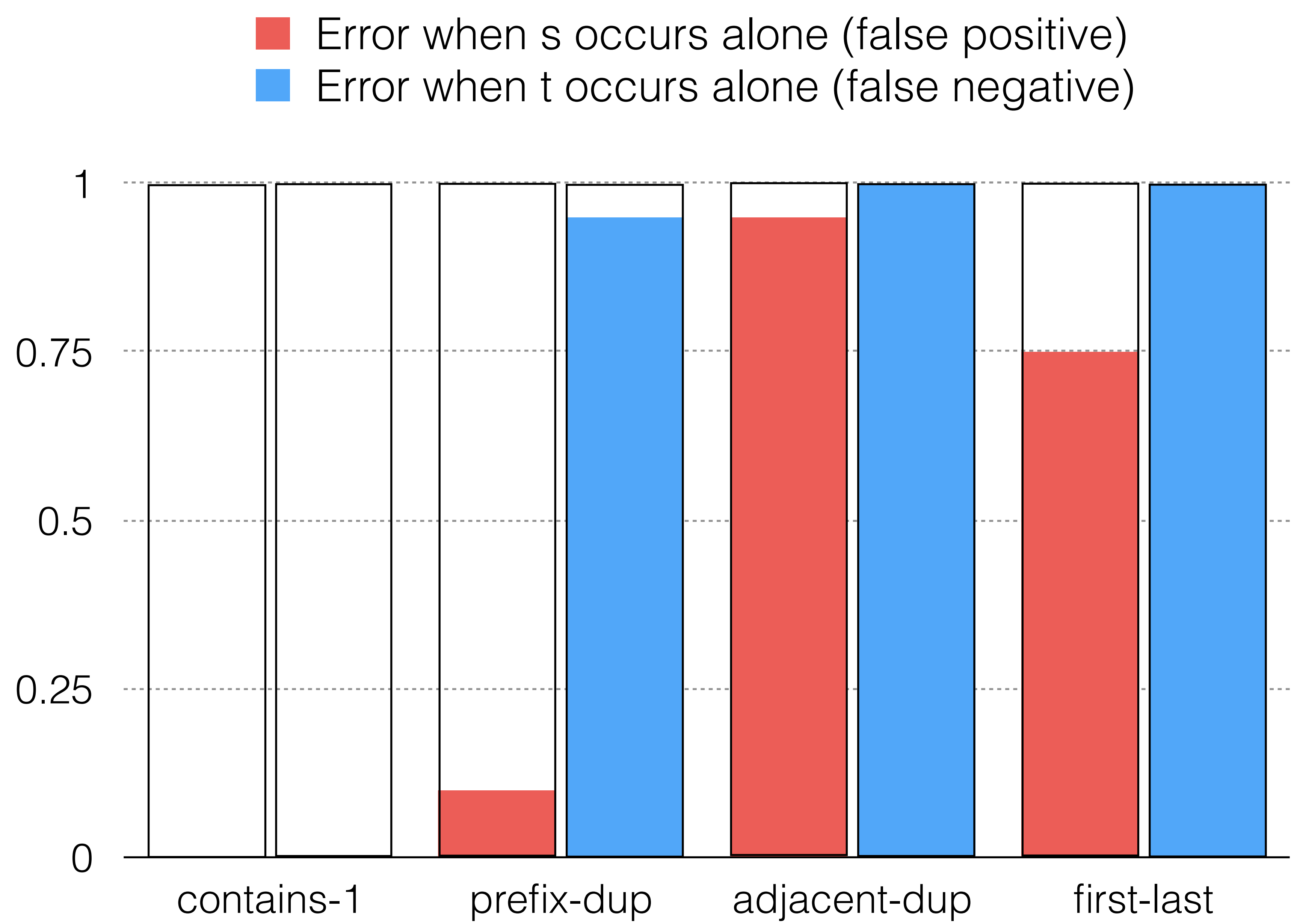
Spurious occurs without target in **0.1%** of training examples



Predicting Inductive Biases of Pretrained Models
Jha, Lovering, Linzen, and Pavlick (2020)

Out-of-Distribution Test Error

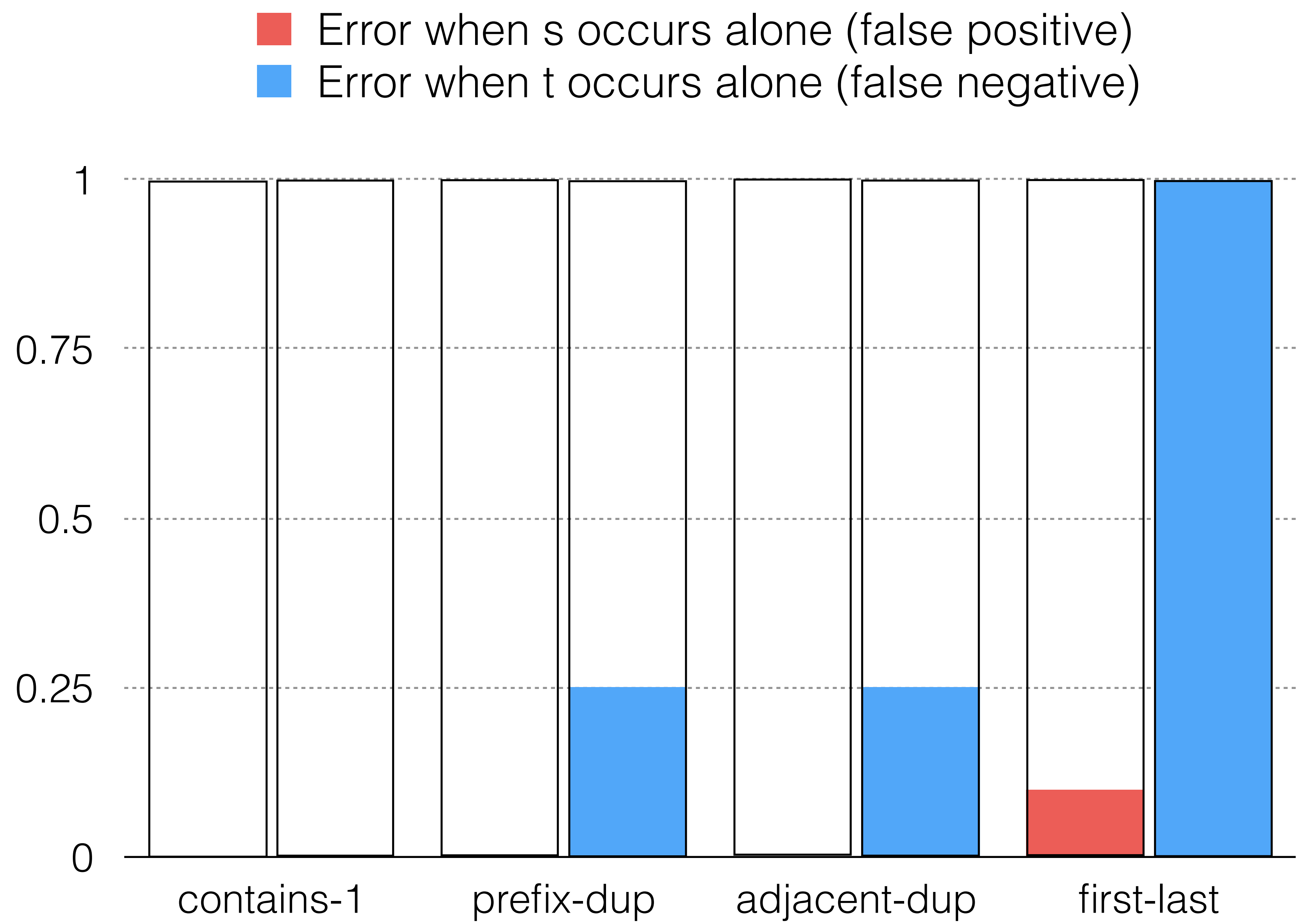
Spurious occurs without target in **10%** of training examples



Predicting Inductive Biases of Pretrained Models
Jha, Lovering, Linzen, and Pavlick (2020)

Out-of-Distribution Test Error

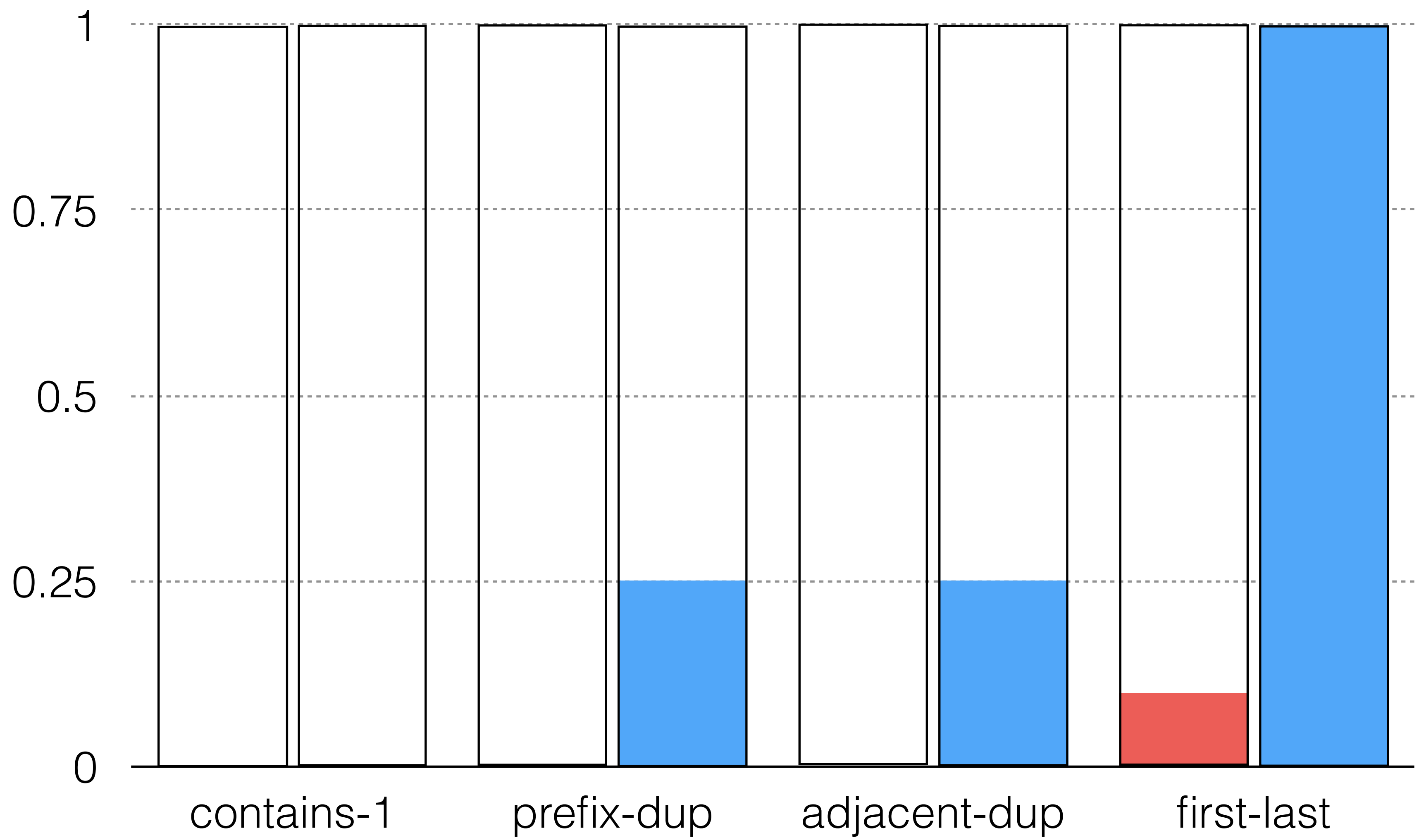
Spurious occurs without target in **50%** of training examples



Out-of-Dist

Different features behave differently given the same training data.

Spurious occurs without target in **50%** of training examples



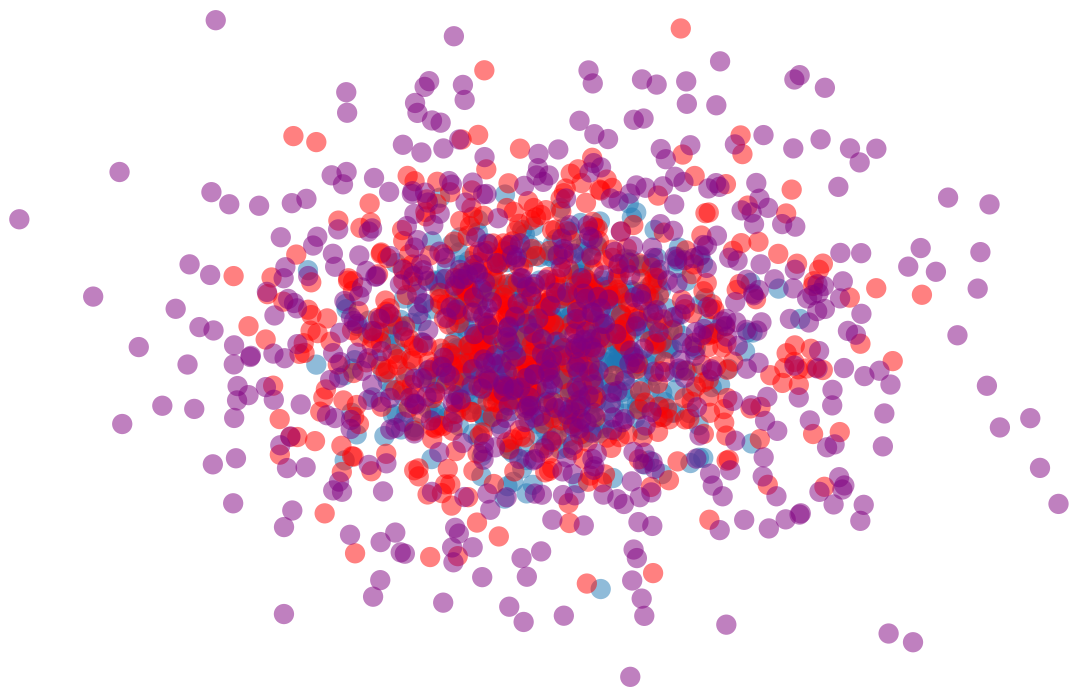
Features differ in how **“hard”** they are to extract

Information-Theoretic Probing with Minimum Description Length. Voita and Titov (2020)

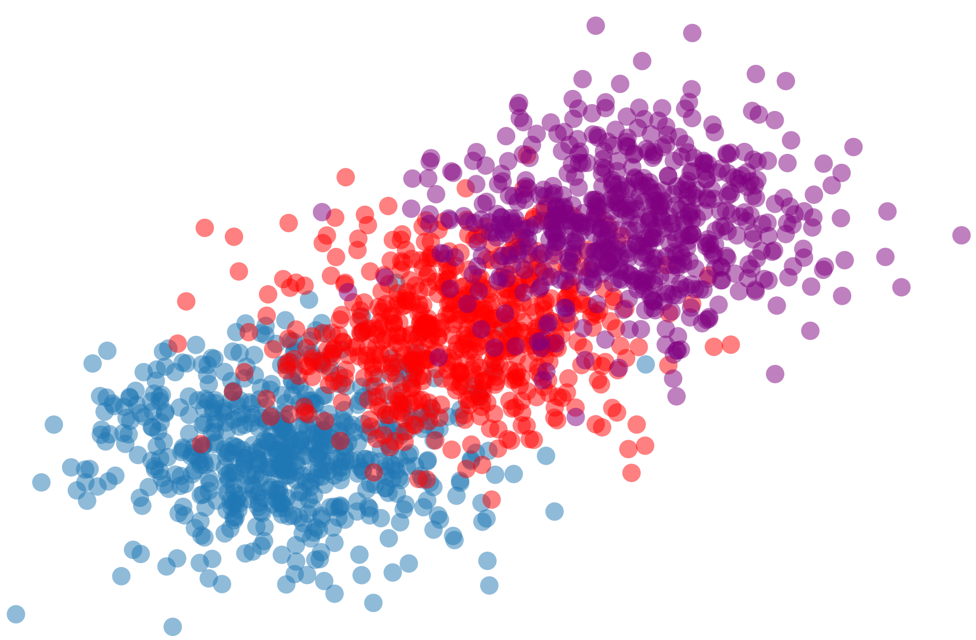
Features differ in how **“hard”** they are to extract

Information-Theoretic Probing with Minimum Description Length. Voita and Titov (2020)

Hard



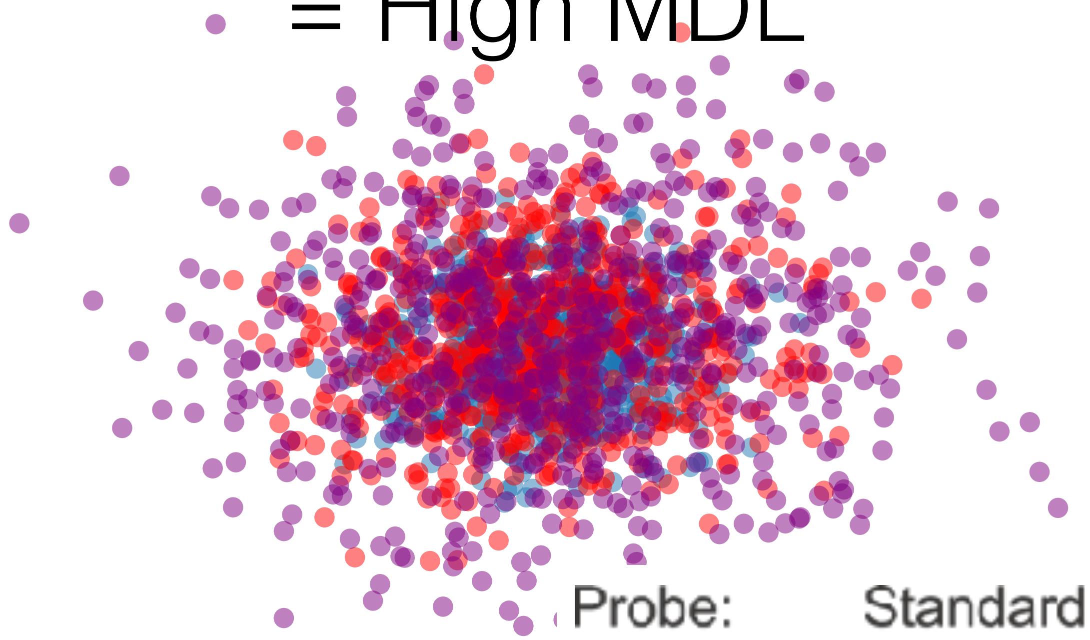
Easy



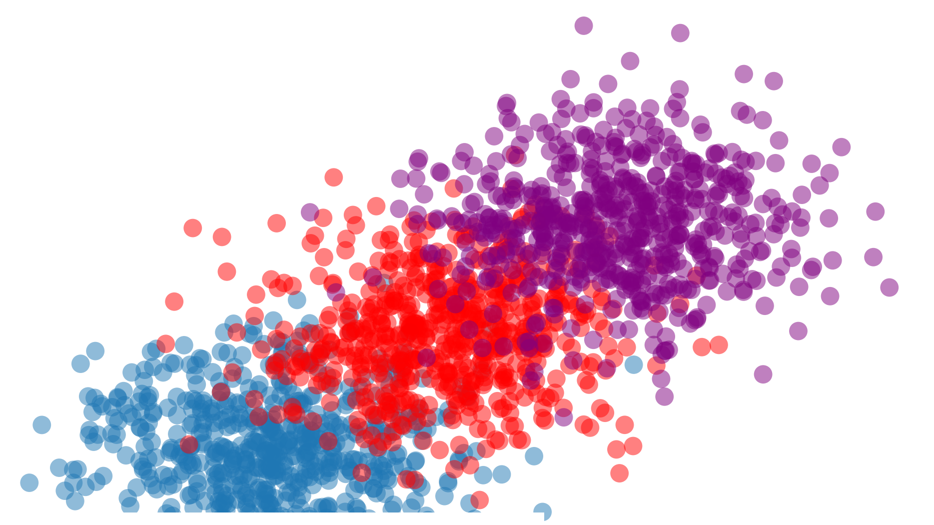
Features differ in how **“hard”** they are to extract

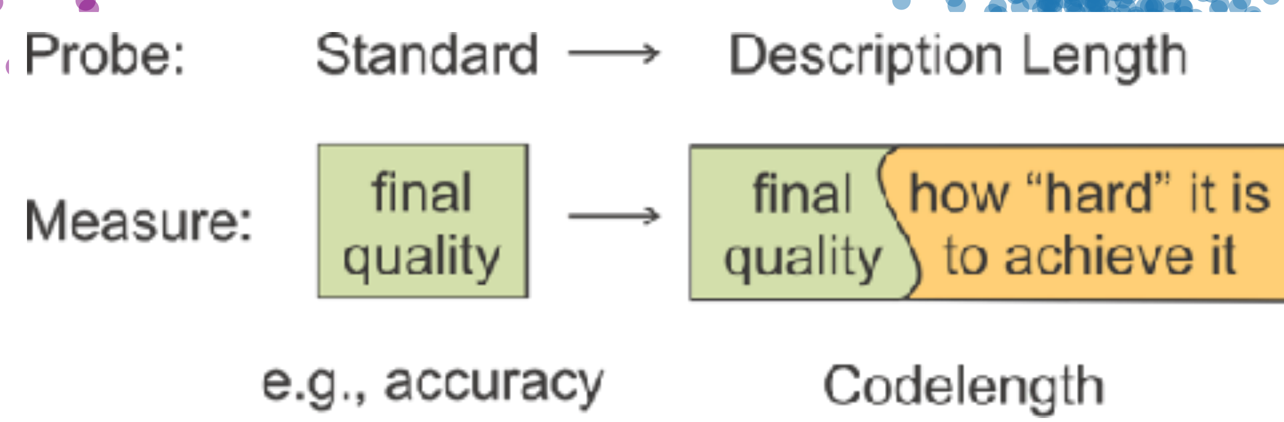
Information-Theoretic Probing with Minimum Description Length. Voita and Titov (2020)

Hard
= High MDL



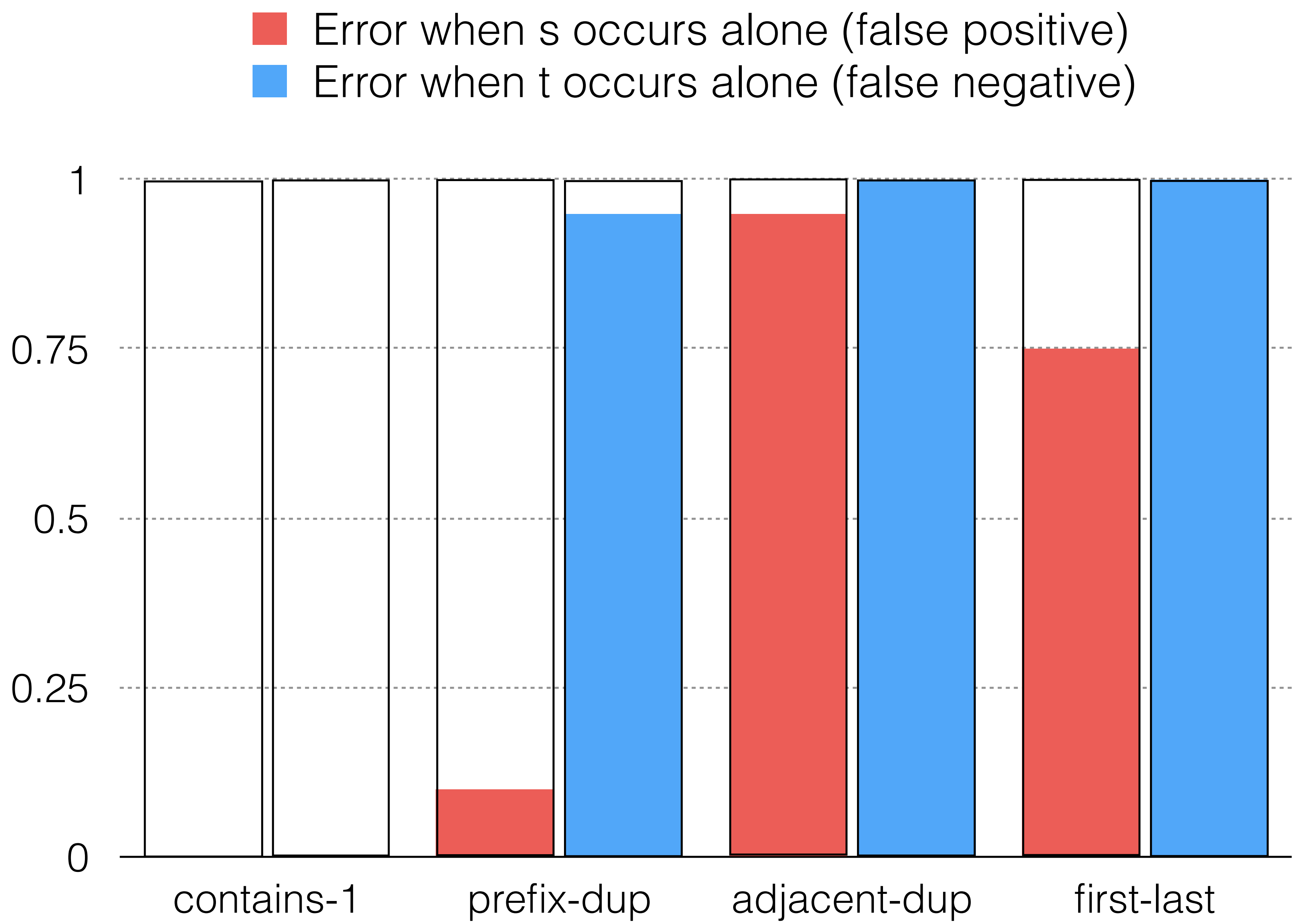
Easy
= Low MDL

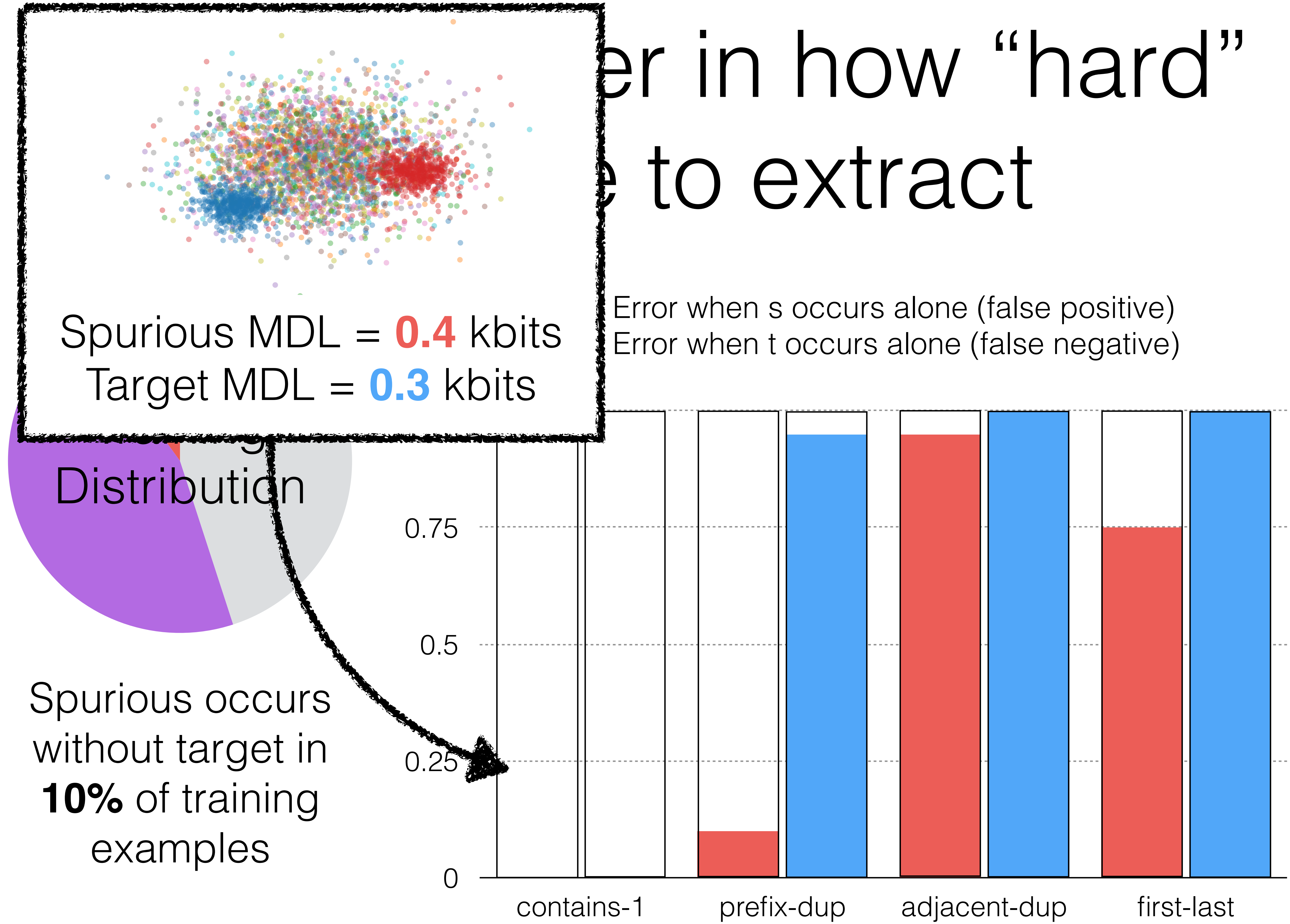


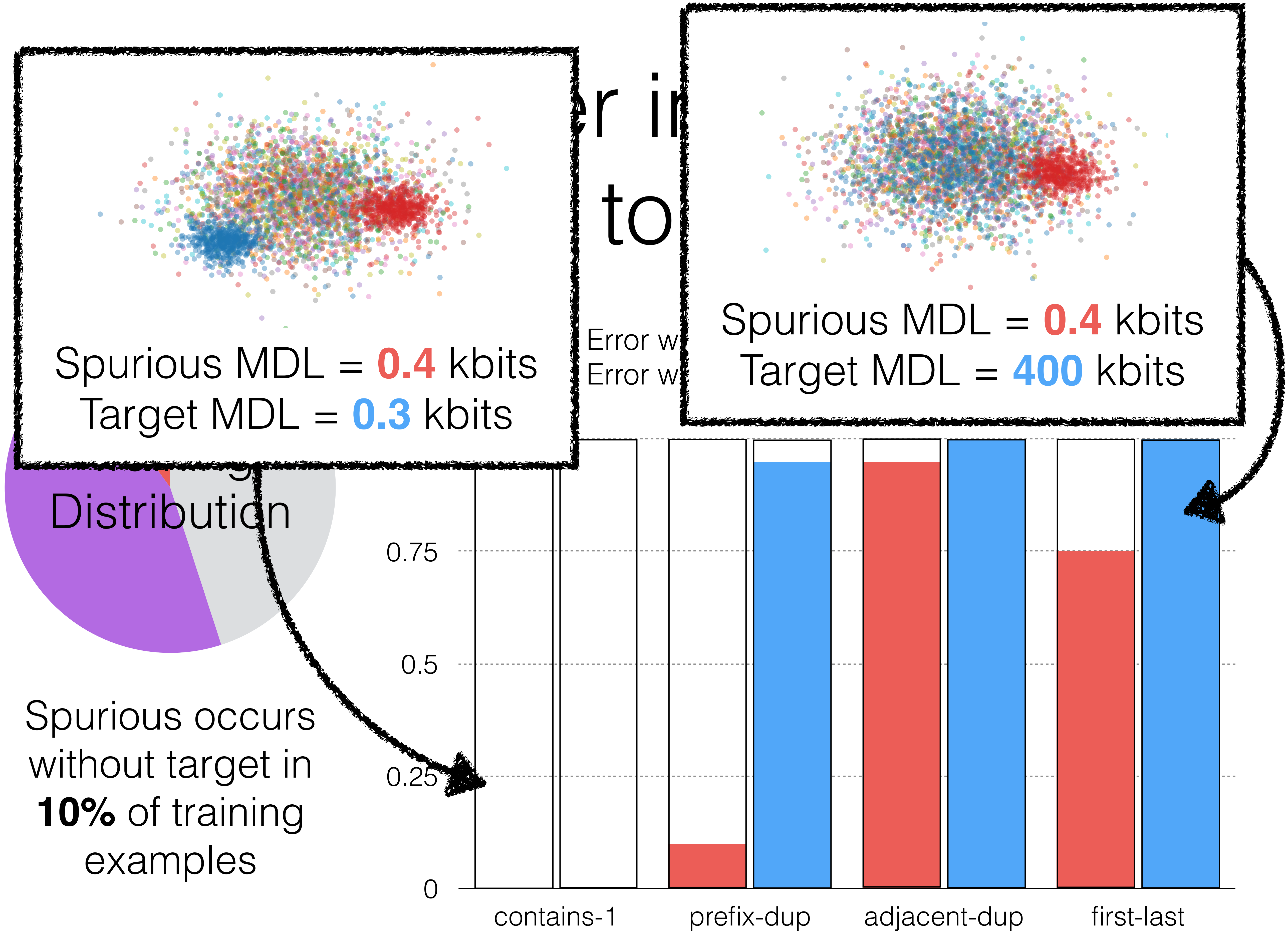


Features differ in how “hard” they are to extract

Spurious occurs without target in **10%** of training examples




er in how “hard”
e to extract
Spurious MDL = 0.4 kbits
Target MDL = 0.3 kbits
Distribution
Spurious occurs without target in 10% of training examples
Error when s occurs alone (false positive)
Error when t occurs alone (false negative)
0.75
0.5
0.25
0
contains-1
prefix-dup
adjacent-dup
first-last



Spurious MDL = 0.4 kbits
Target MDL = 0.3 kbits
Spurious MDL = 0.4 kbits
Target MDL = 400 kbits
Distribution
Spurious occurs without target in **10%** of training examples
Error w
Error w
0.75
0.5
0.25
0
contains-1
prefix-dup
adjacent-dup
first-last

Hypothesis

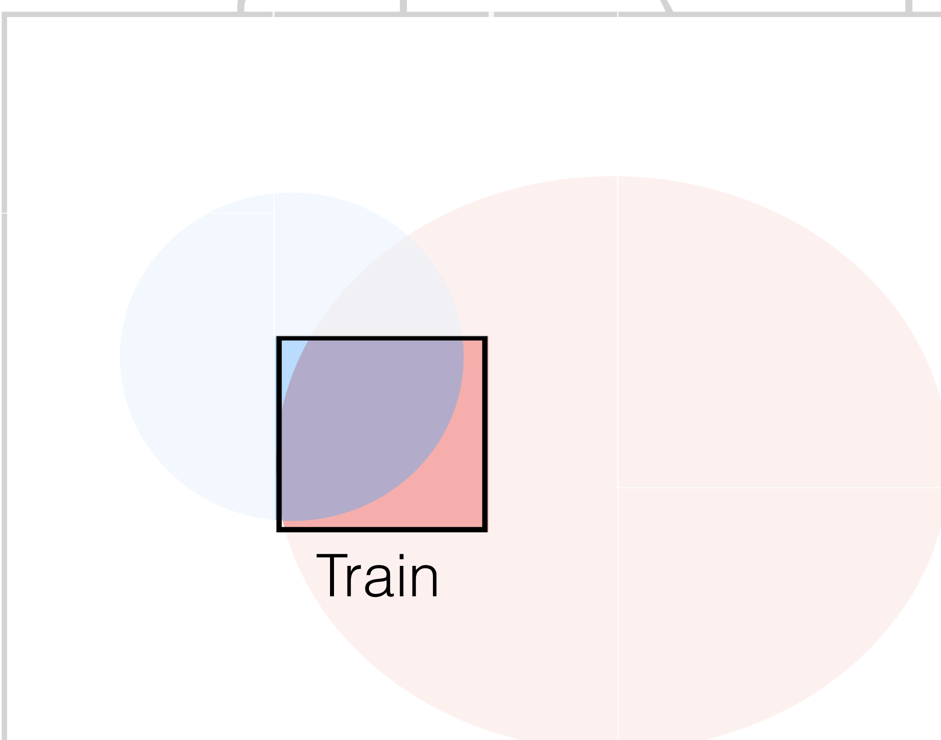
A fine-tuned model's use of a feature (the "target") is a function of both the difficulty of extracting the feature (relative to competing "spurious" features) and the training evidence against the competing spurious features.

Hypothesis

A fine-tuned model's use of a feature (the "target") is a function of both the difficulty of extracting the feature (relative to competing "spurious"

the **training evidence**

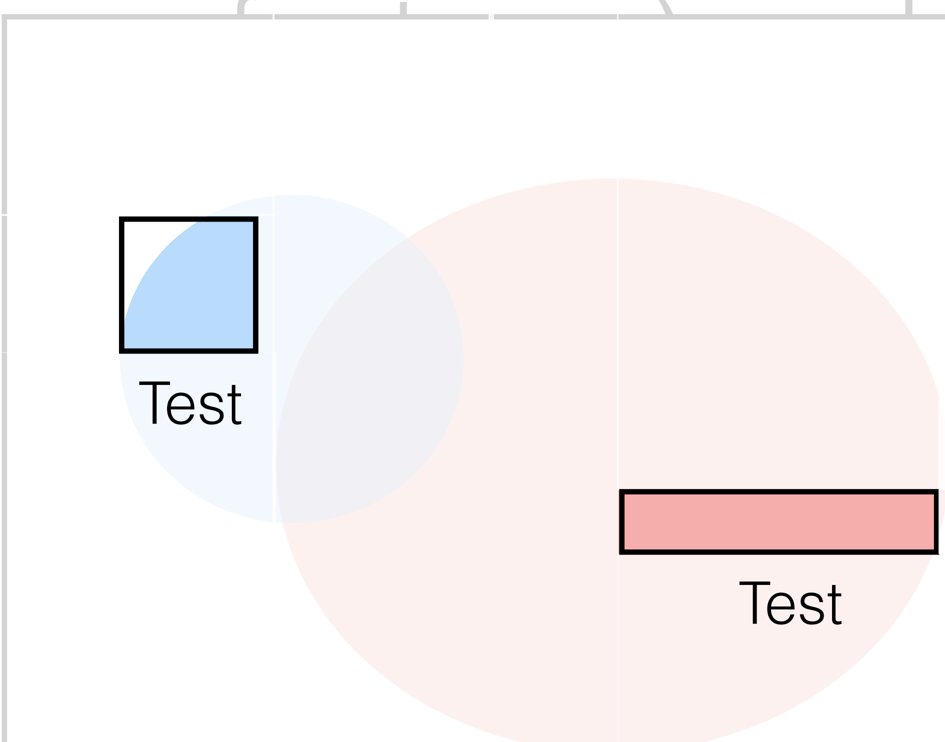
peting spurious features.



(Lack of) co-occurrence between spurious and target during training

Hypothesis

A fine-tuned model's **use of a feature** (the "target") is a function of both the difficulty of extracting the feature (relative to competing "spurious" the **training evidence** peting spurious features.



Performance on out-of-distribution test set

Hypothesis

A fine-tuned model's **use of a feature** (the "target") is a function of both the **difficulty of extracting the feature** (relative to competing "spurious" features) and the **training evidence** against the competing spurious features.

$$\frac{\texttt{MDL of spurious}}{\texttt{MDL of target}}$$

Higher -> Target is comparatively easier extract

Experimental Set Up

Experimental Set Up

Task: Sentence Acceptability

`The piano teachers see the handyman.`

✔

Experimental Set Up

Task: Sentence Acceptability

`The piano teachers sees the handyman.`

✗

Experimental Set Up

Task: Sentence Acceptability

Target Feature: Subject-Verb Agreement

`The piano teachers of the lawyer see the handyman.`

Experimental Set Up

Task: Sentence Acceptability

Target Feature: Subject-Verb Agreement

Spurious Feature #1: Lexical Item

`Often, the piano teachers of the lawyer see the handyman.`

Experimental Set Up

Task: Sentence Acceptability

Target Feature: Subject-Verb Agreement

Spurious Feature #2: Sentence Length

`The piano teachers of the lawyer who works in the city across the river see the handyman.`

Experimental Set Up

Task: Sentence Acceptability

Target Feature: Subject-Verb Agreement

Spurious Feature #3: Plural Nouns

`The piano teachers of the lawyers see the handyman.`

Experimental Set Up

Task: Sentence Acceptability

Target Feature: Subject-Verb Agreement

Spurious Feature #4: Closest Noun Agreement

`The piano teachers of the lawyers see the handyman.`

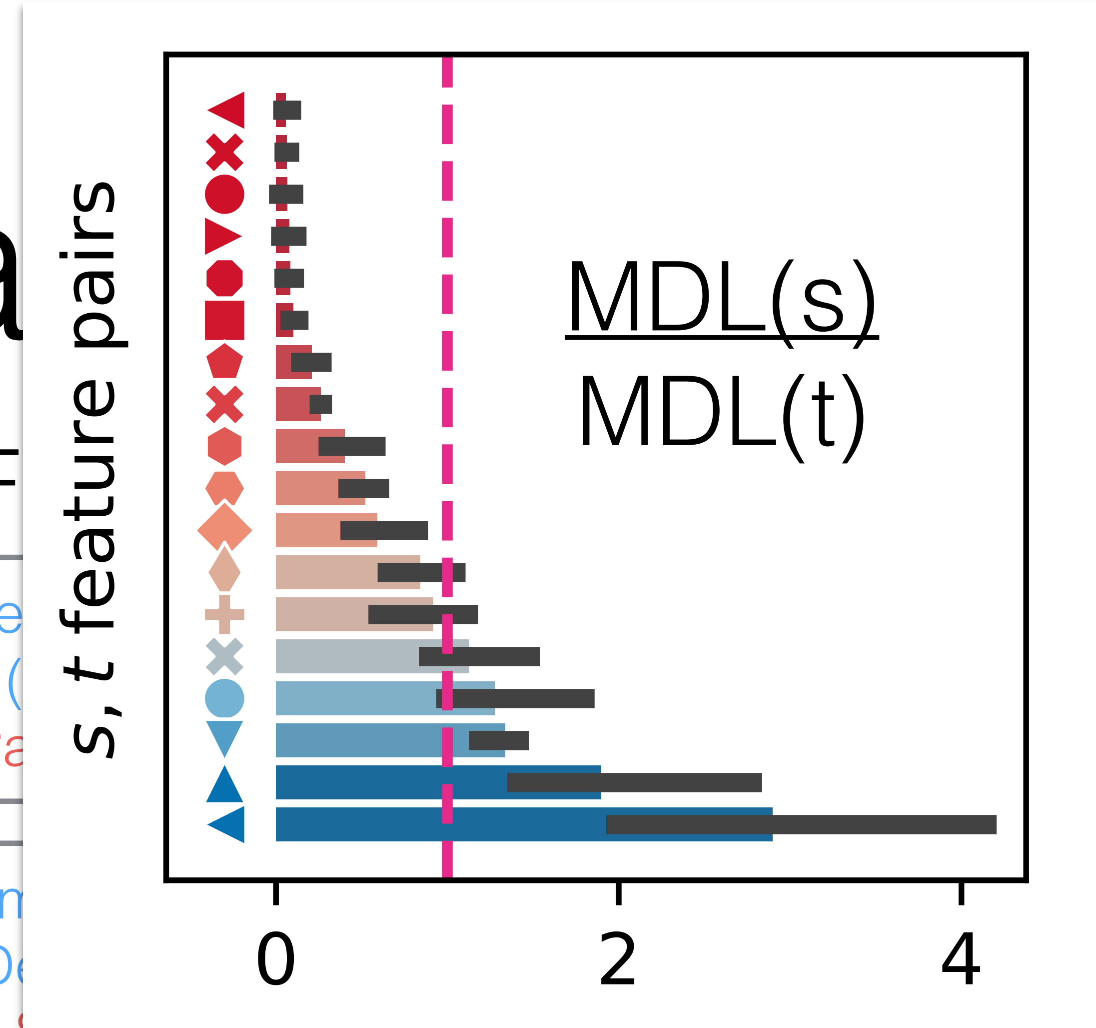
Experimental Set Up

20 Target-Spurious Feature Pairs

T: Simple Subj-Verb Agr. S: Lexical	T: Subj-Verb Agr. w/ Attractors S: Plural Noun	T: Negative Polarity Item (NPI) Licensing S: Past Tense Verb	T: Hard Filler-Gap Dependency S: Lexical
T: Simple Subj-Verb Agr. S: Plural Noun	T: Subj-Verb Agr. w/ Attractors S: Closest Noun	T: Simple Filler-Gap Dependency S: Lexical	T: Hard Filler-Gap Dependency S: Length
T: Simple Subj-Verb Agr. S: Closest Noun	T: Negative Polarity Item (NPI) Licensing S: Lexical	T: Simple Filler-Gap Dependency S: Length	T: Hard Filler-Gap Dependency S: Plural Noun
T: Subj-Verb Agr. w/ Attractors S: Lexical	T: Negative Polarity Item (NPI) Licensing S: Length	T: Simple Filler-Gap Dependency S: Plural Noun	T: Hard Filler-Gap Dependency S: Past Tense Verb
T: Subj-Verb Agr. w/ Attractors S: Length	T: Negative Polarity Item (NPI) Licensing S: Plural Noun	T: Simple Filler-Gap Dependency S: Past Tense Verb	T: Hard Filler-Gap Dependency S: None

Experimenta

20 Target-Spurious F

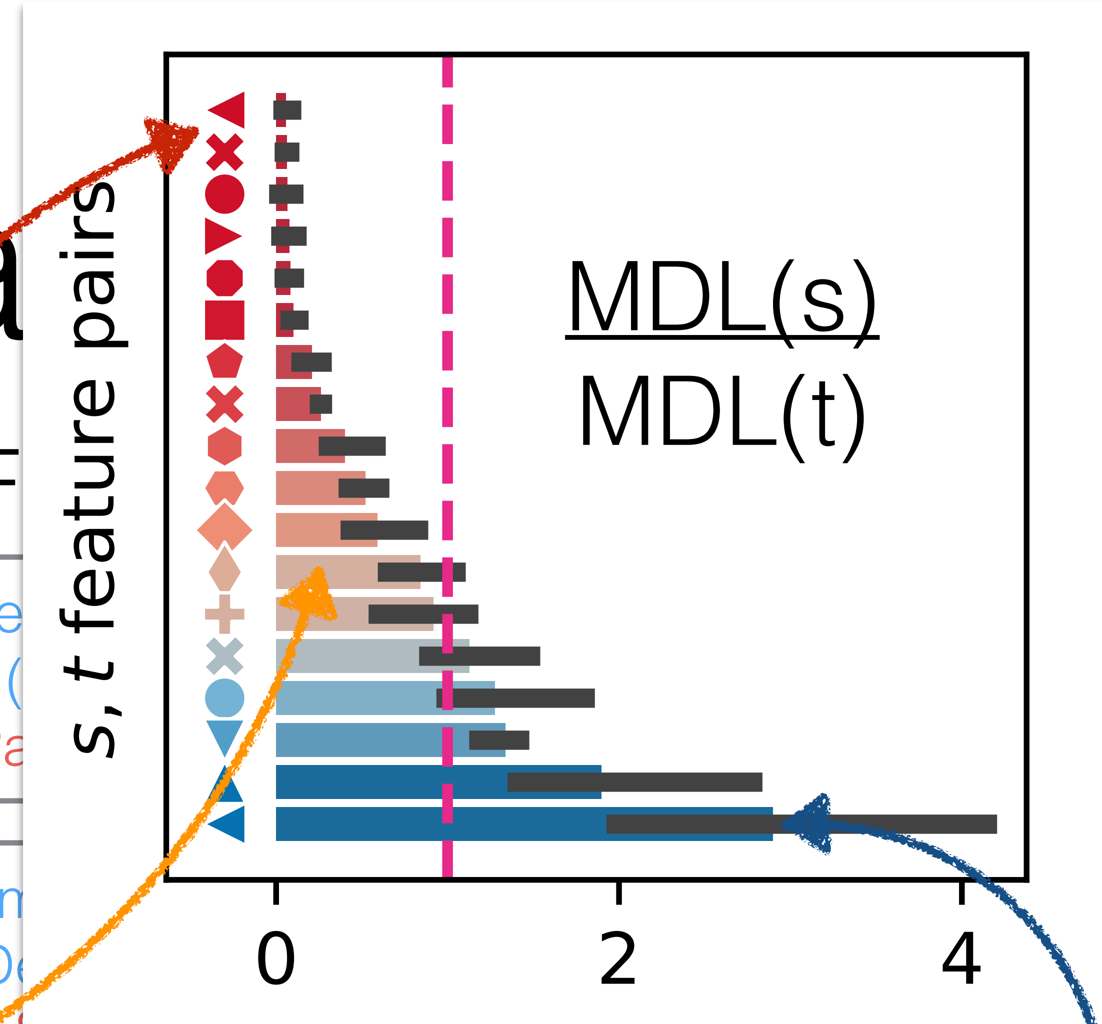


T: Simple Subj-Verb Agr. S: Lexical	T: Subj-Verb Agr. w/ Attractors S: Plural Noun	T: Ne Item (S: Pa	
T: Simple Subj-Verb Agr. S: Plural Noun	T: Subj-Verb Agr. w/ Attractors S: Closest Noun	T: Sim De S: Lexical	S: Length
T: Simple Subj-Verb Agr. S: Closest Noun	T: Negative Polarity Item (NPI) Licensing S: Lexical	T: Simple Filler-Gap Dependency S: Length	T: Hard Filler-Gap Dependency S: Plural Noun
T: Subj-Verb Agr. w/ Attractors S: Lexical	T: Negative Polarity Item (NPI) Licensing S: Length	T: Simple Filler-Gap Dependency S: Plural Noun	T: Hard Filler-Gap Dependency S: Past Tense Verb
T: Subj-Verb Agr. w/ Attractors S: Length	T: Negative Polarity Item (NPI) Licensing S: Plural Noun	T: Simple Filler-Gap Dependency S: Past Tense Verb	T: Hard Filler-Gap Dependency S: None

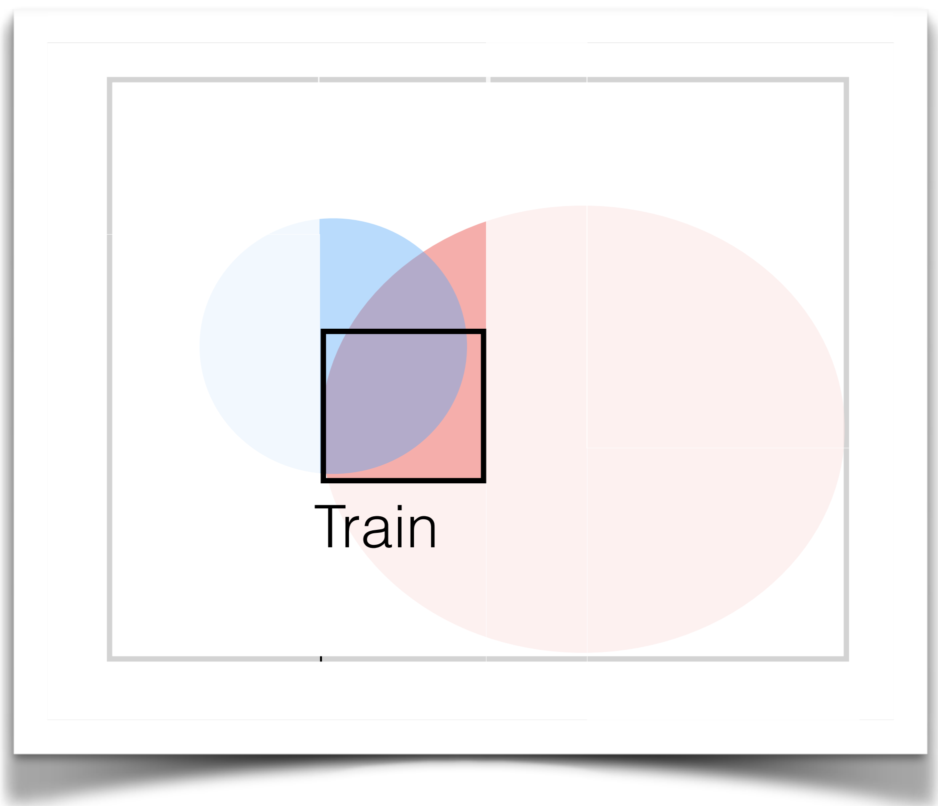
Experimenta

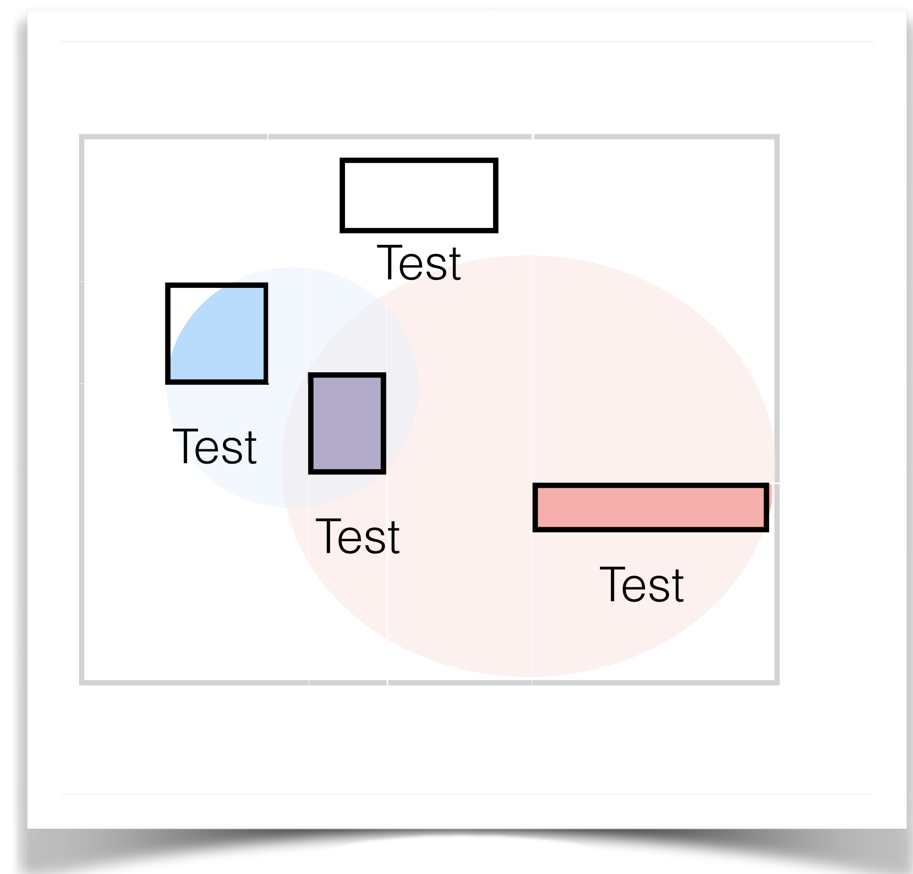
20 Target-Spurious F

T: Simple Subj-Verb Agr. S: Lexical	T: Subj-Verb Agr. w/ Attractors S: Plural Noun	T: Ne Item (S: Pa	
T: Simple Subj-Verb Agr. S: Plural Noun	T: Subj-Verb Agr. w/ Attractors S: Closest Noun	T: Sim De S: Lexical	S: Length
T: Simple Subj-Verb Agr. S: Closest Noun	T: Negative Polarity Item (NPI) Licensing S: Lexical	T: Simple Filler-Gap Dependency S: Length	T: Hard Filler-Gap Dependency S: Plural Noun
T: Subj-Verb Agr. w/ Attractors S: Lexical	T: Negative Polarity Item (NPI) Licensing S: Length	T: Simple Filler-Gap Dependency S: Plural Noun	T: Hard Filler-Gap Dependency S: Past Tense Verb
T: Subj-Verb Agr. w/ Attractors S: Length	T: Negative Polarity Item (NPI) Licensing S: Plural Noun	T: Simple Filler-Gap Dependency S: Past Tense Verb	T: Hard Filler-Gap Dependency S: None



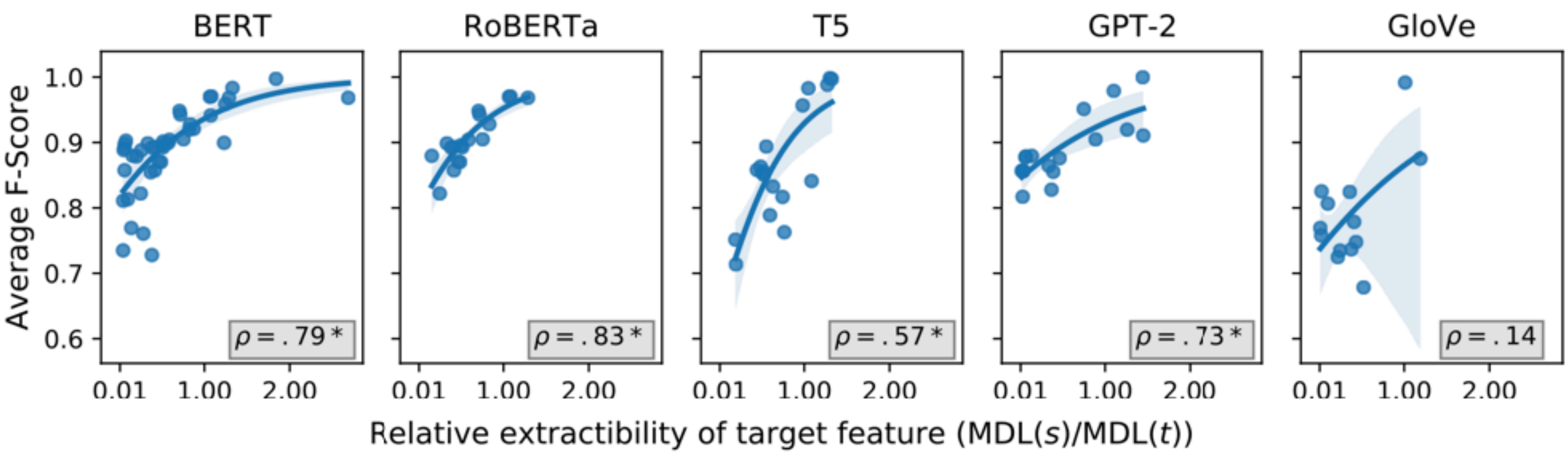
Results



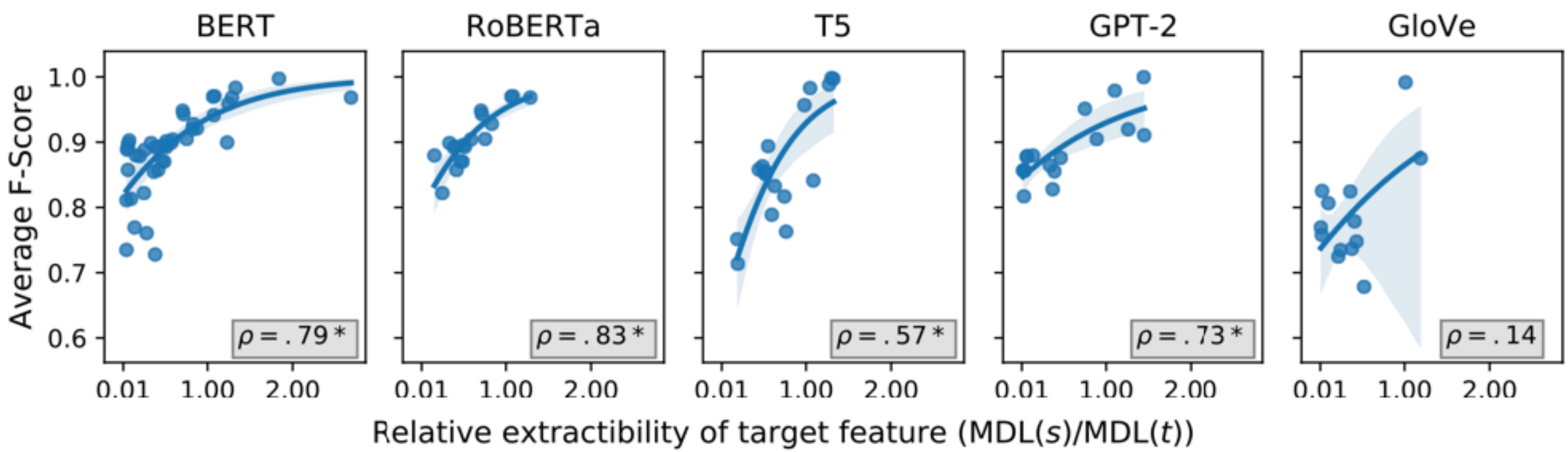


"Average F Score"

Results

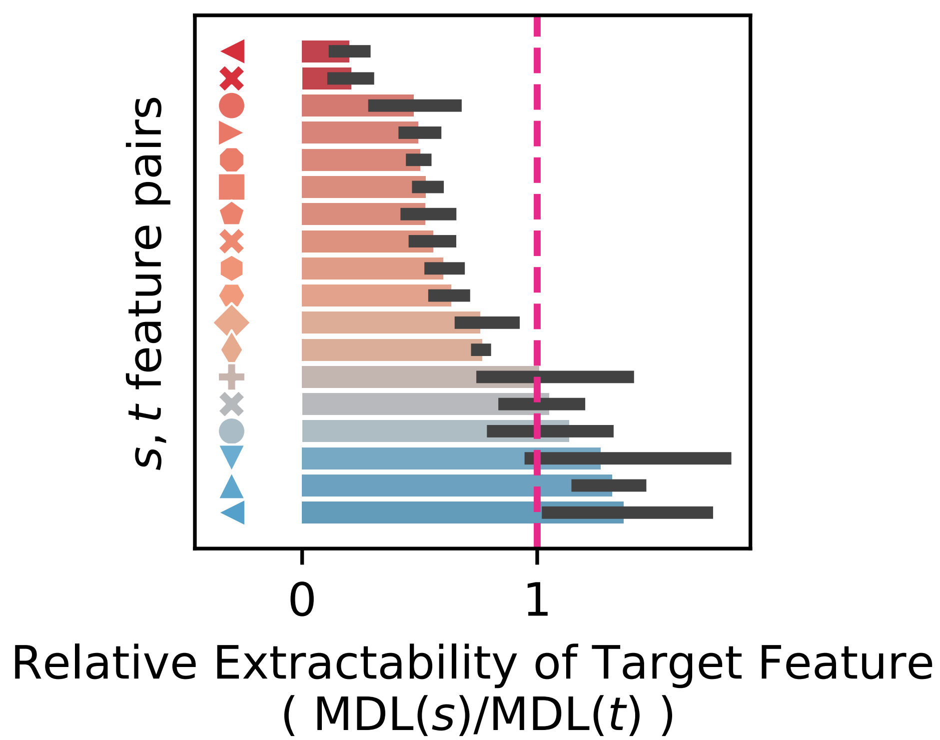

BERT
RoBERTa
T5
GPT-2
GloVe
Average F-Score
1.0
0.9
0.8
0.7
0.6
ρ = .79*
ρ = .83*
ρ = .57*
ρ = .73*
ρ = .14
0.01
1.00
2.00
Relative extractibility of target feature (MDL(s)/MDL(t))

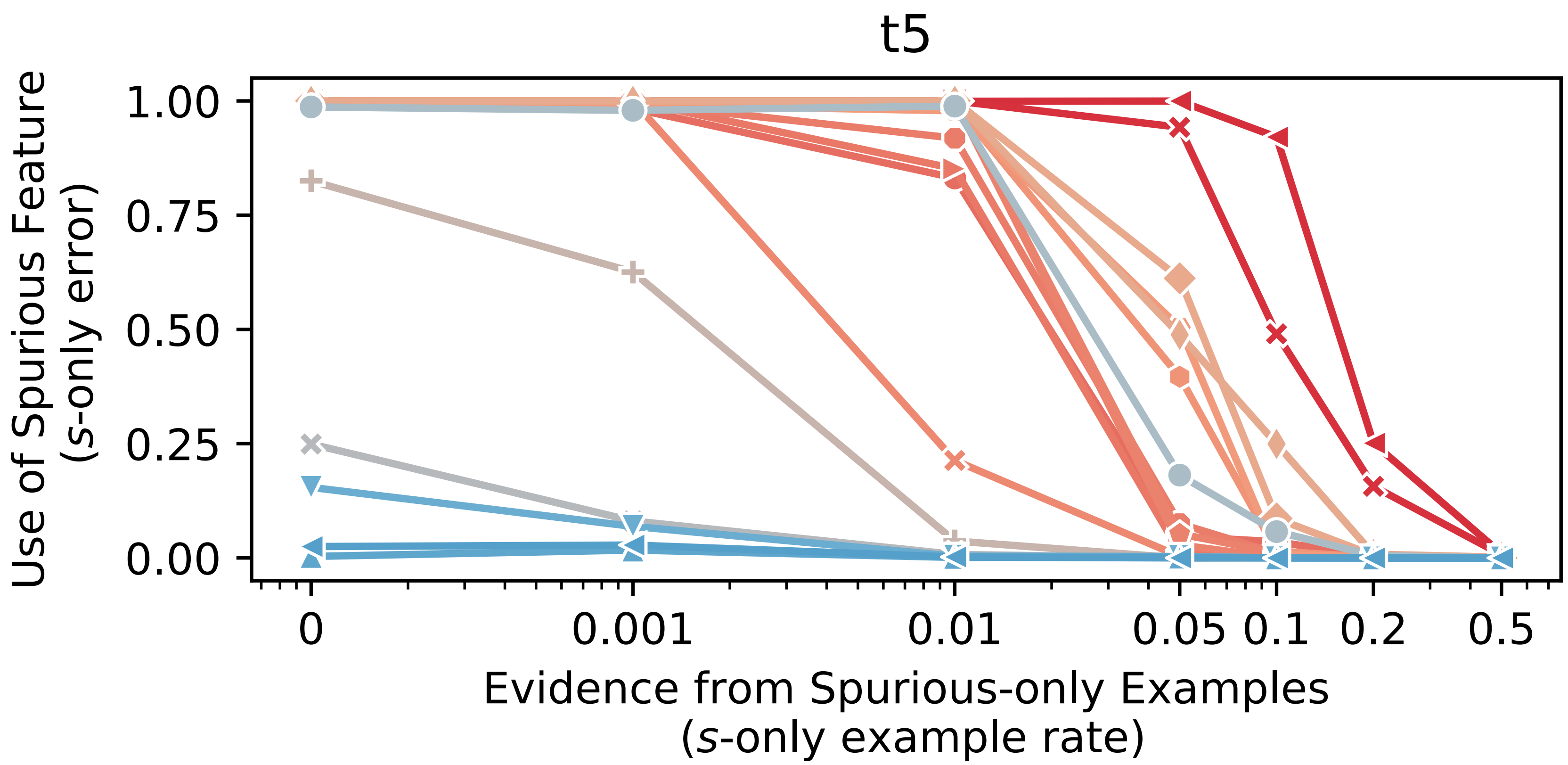
Results



The easier target is to extract relative to spurious, the more likely the model is to use the target feature.

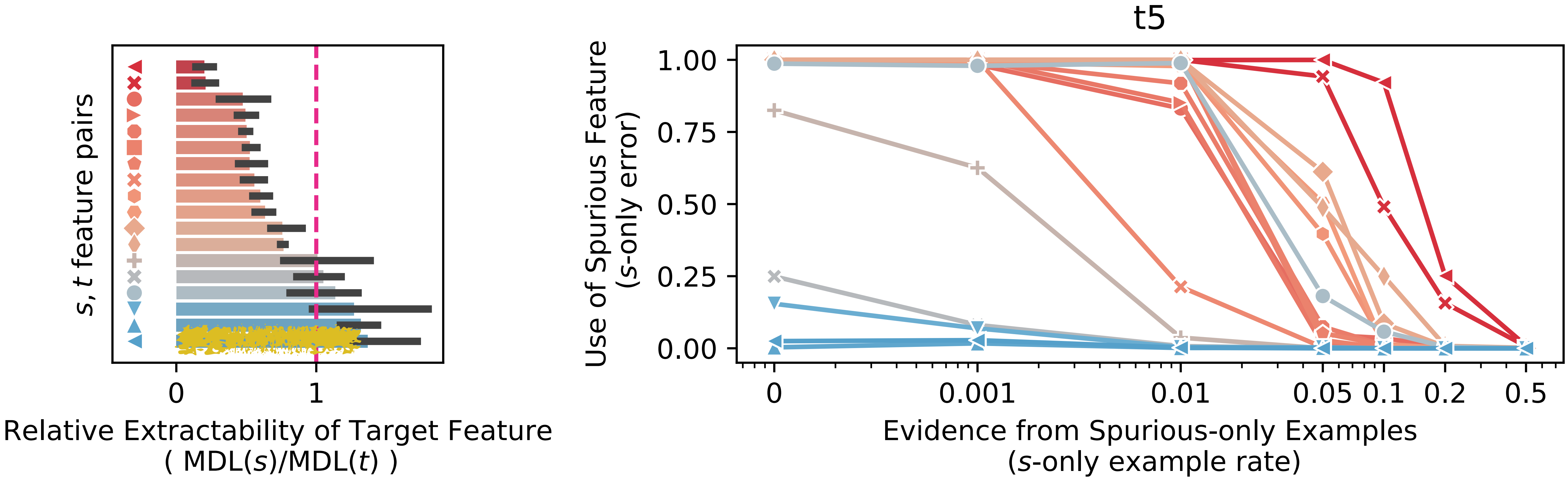
Results





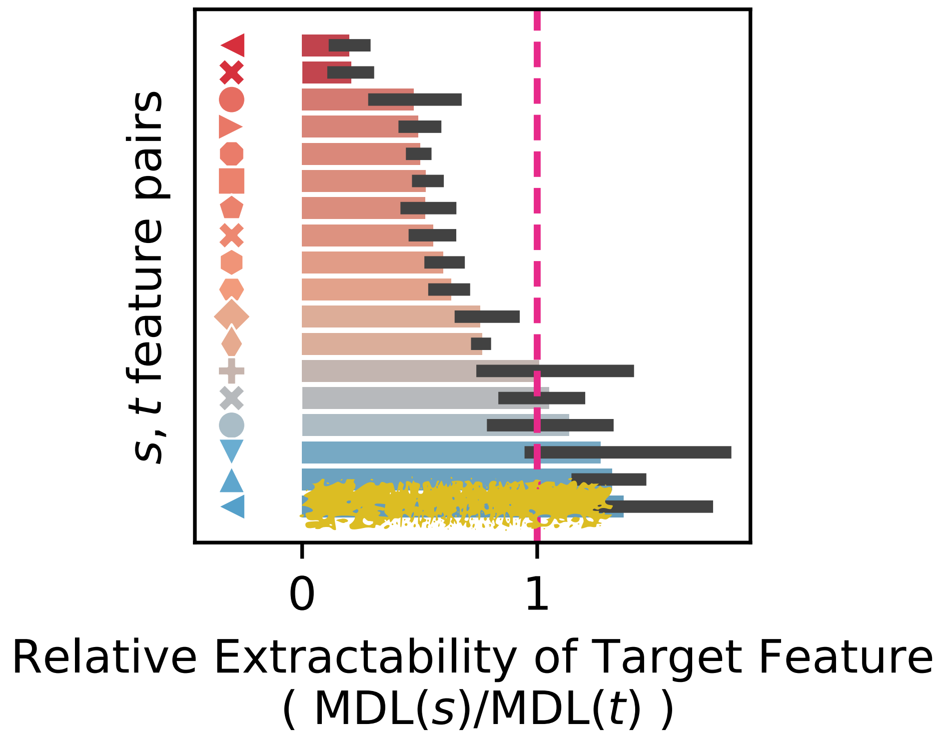
The easier target is to extract relative to spurious, the less sensitive the model is to priors in the training data.

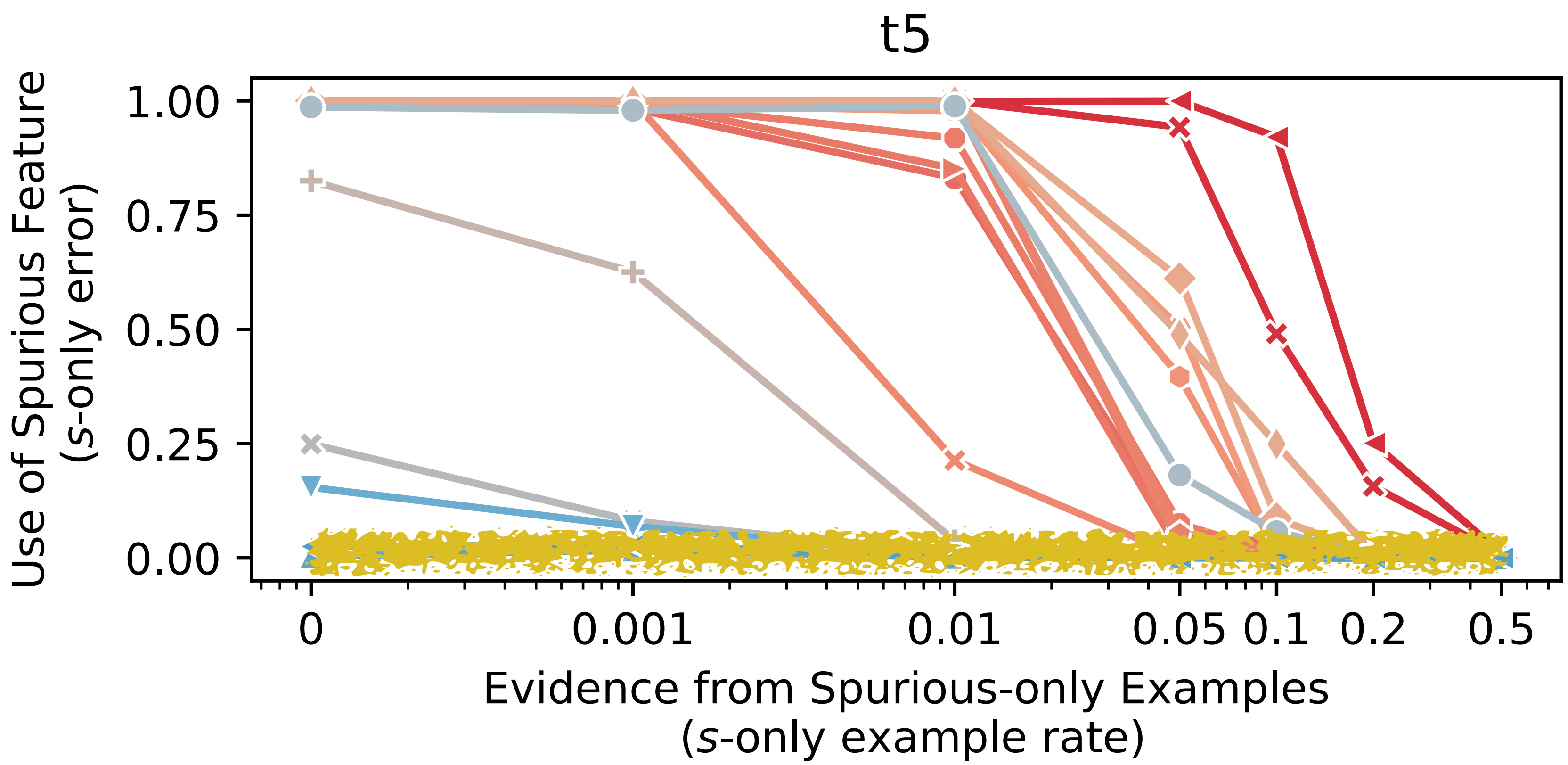
Results



When target is much easier to extract than spurious...

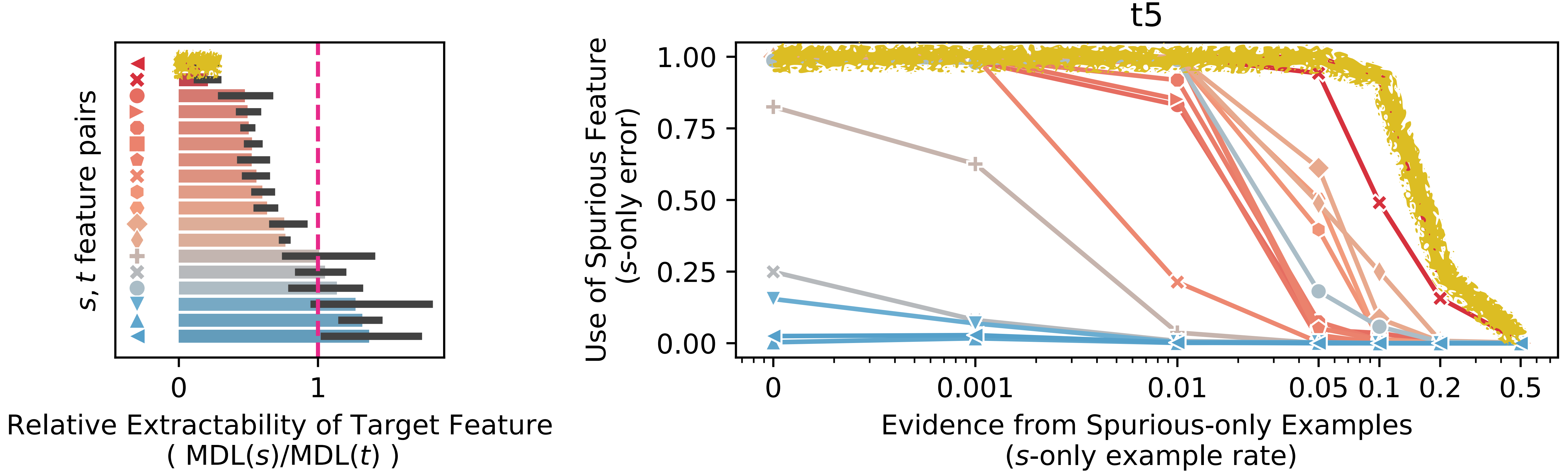
Results





When target is much easier to extract than spurious...model learns the right thing despite no training incentive to do so.

Results



When target is much harder to extract than spurious...model requires substantial training incentive (e.g., 5% of training examples).

Do NNs apply systematic rules?

(NLP/Syntax)

Frequency Effects on Syntactic Rule Learning in Transformers
Wei et al (under review)

Do NNs *have* symbolic concepts?

(Computer Vision)

✅ `is_grounded`

✅ `is_token_of_type`

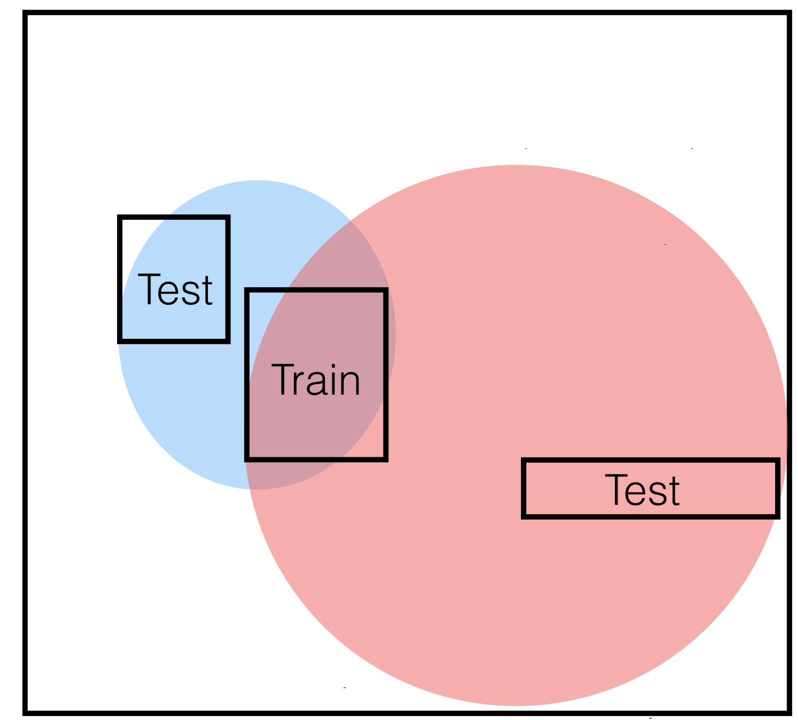
⚠️ `is_contxt_independent`

⛔ `is_causal`

Unit Testing for Concepts in Neural Networks
Lovering & Pavlick (in progress)

Why don't NNs *use* symbols and rules, even if they can?

(Toy, NLP/Syntax)



Predicting Inductive Biases of Pretrained Models
Lovering et al (ICLR, 2021)

Do NNs apply systematic rules?

(NLP/Syntax)

Frequency Effects on Syntactic Rule Learning in Transformers
Wei et al (under review)

Do NNs *have* symbolic concepts?

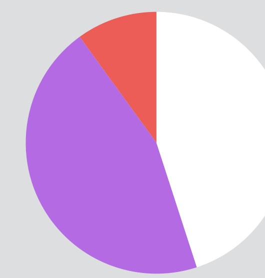
(Computer Vision)

- ✅ `is_grounded`
- ✅ `is_token_of_type`
- ⚠️ `is_contxt_independent`
- ⛔ `is_causal`

Unit Testing for Concepts in Neural Networks
Lovering & Pavlick (in progress)

Why don't NNs *use* symbols and rules, even if they can?

(Toy, NLP/Syntax)



Predicting Inductive Biases of Pretrained Models
Lovering et al (ICLR, 2021)

Do NNs apply systematic rules? (NLP/Syntax)

Do NNs *have* symbolic concepts? (Computer Vision)

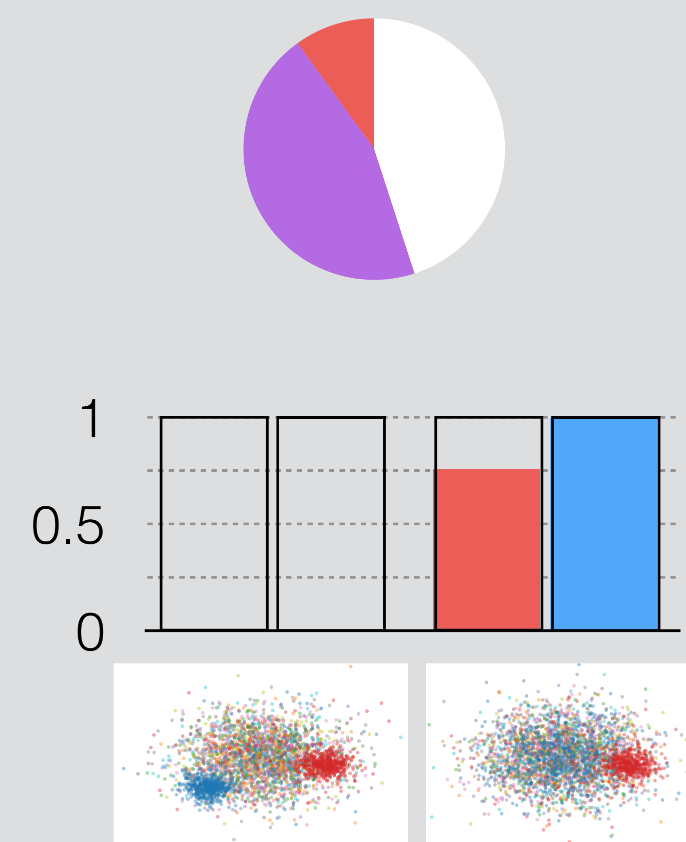
- Models don't necessarily solve the task the best way…**even when they are capable of doing so**

Frequency Effects on Syntactic Rule Learning in Transformers
Wei et al (under review)

Unit Testing for Concepts in Neural Networks
Lovering & Pavlick (in progress)

Why don't NNs *use* symbols and rules, even if they can?

(Toy, NLP/Syntax)



Predicting Inductive Biases of Pretrained Models
Lovering et al (ICLR, 2021)

Do NNs apply systematic rules?

(NLP/Syntax)

Frequency Effects on Syntactic Rule Learning in Transformers
Wei et al (under review)

Do NNs *have* symbolic concepts?

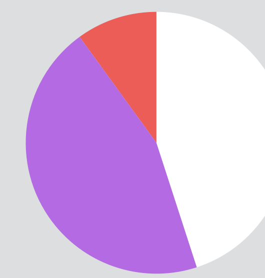
(Computer Vision)

Unit Testing for Concepts in Neural Networks
Lovering & Pavlick (in progress)

- Models don't necessarily solve the task the best way…**even when they are capable of doing so**
- Models sometimes struggled to overcome **strong training data priors**

Why don't NNs *use* symbols and rules, even if they can?

(Toy, NLP/Syntax)



Predicting Inductive Biases of Pretrained Models
Lovering et al (ICLR, 2021)

Do NNs apply systematic rules?

(NLP/Syntax)

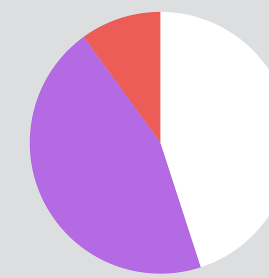
Do NNs *have* symbolic concepts?

(Computer Vision)

- Models don't necessarily solve the task the best way…**even when they are capable of doing so**
- Models sometimes struggled to overcome **strong training data priors**
- But, when feature representations are sufficiently well encoded, models show **correct inductive biases** and generalize well **despite little/no training incentive** to do so

Why don't NNs *use* symbols and rules, even if they can?

(Toy, NLP/Syntax)

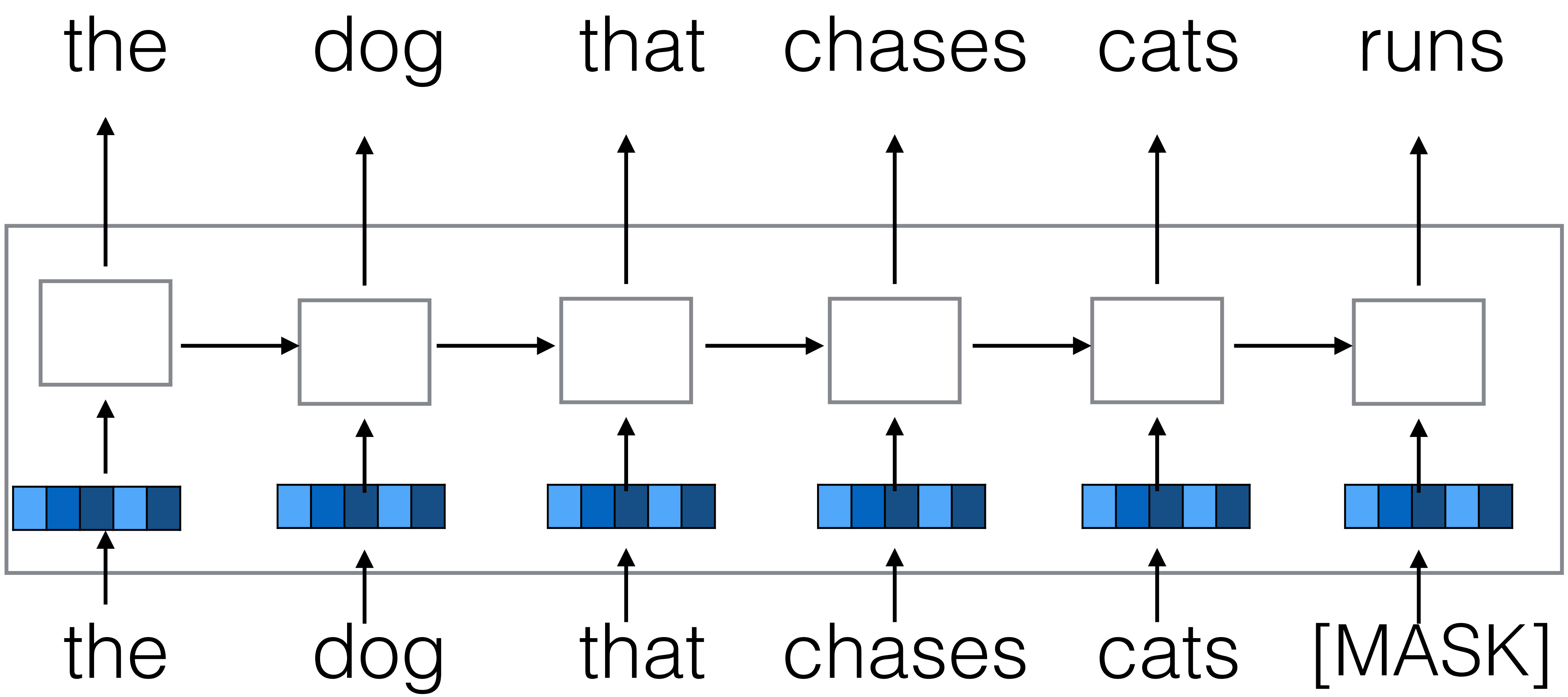


Predicting Inductive Biases of Pretrained Models
Lovering et al (ICLR, 2021)

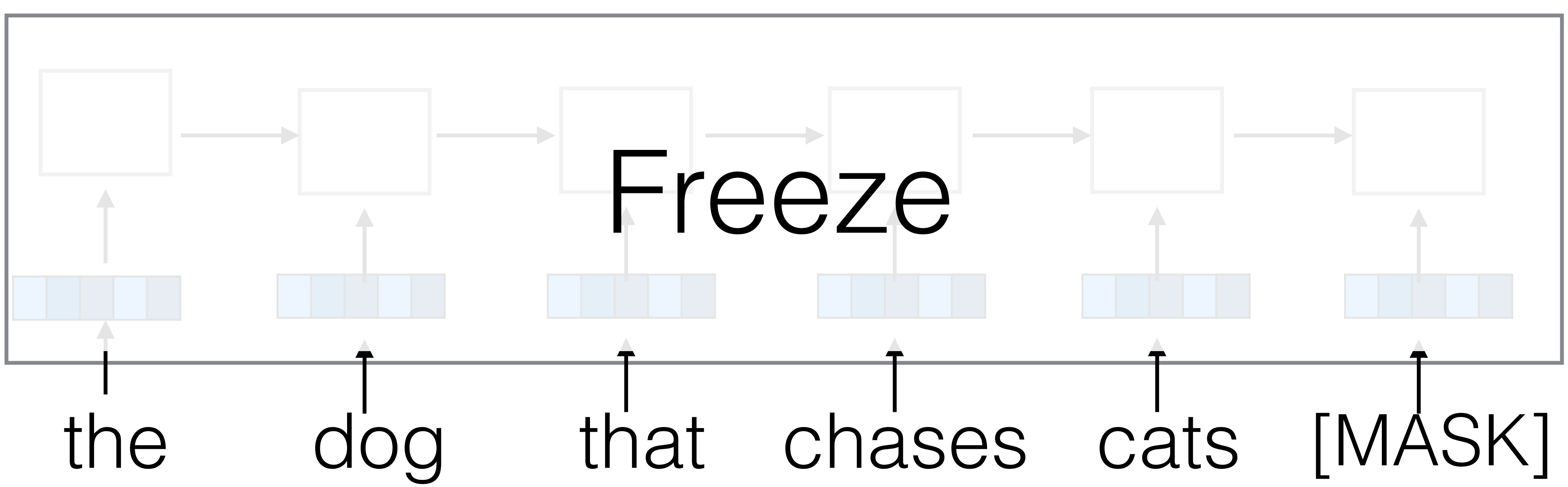
Frequency Effects on Syntactic Rule Learning in Transformers
Wei et al (under review)

Unit Testing for Concepts in Neural Networks
Lovering & Pavlick (in progress)

Predicting Agreement Features



Predicting Agreement Features



Predicting Agreement Features