Language Understanding and
Representation Laboratory

No One Metric is Enough!

Combining Evaluation Techniques to Uncover Latent Structure

Ellie Pavlick, Challenges of Compositionality Workshop, June 30 2022

What does it mean to be “compositional”?

Candidate Definitions

What does it mean to be “compositional”?

Candidate Definitions

The meaning of a sentence is a function of the meanings of the words and the way in which they are combined. (Partee, 1995)

More lenient ⟷ More stringent

What does it mean to be "compositional"?

Candidate Definitions

The meaning of a sentence is a function of the meanings of the words and the way in which they are combined. (Partee, 1995)

More lenient

More stringent

Neural nets meet this definition by construction.

$$y_t = f(W_x x_t + W_h h_{t-1})$$

What does it mean to be “compositional”?

Candidate Definitions

The meaning of a sentence is a function of the meanings of the words and the way in which they are combined.
(Partee, 1995)

More lenient ⟷ More stringent

Neural nets meet this definition by construction.

$$y_t = f(W_x x_t + W_h h_{t-1})$$

Not an interesting intellectual debate.

What does it mean to be "compositional"?

Candidate Definitions

The meaning of a sentence is a function of the meanings of the words and the way in which they are combined. (Partee, 1995)

"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)

More lenient ⟷ More stringent

What does it mean to be "compositional"?

Candidate Definitions

The meaning of a sentence is a function of the meanings of the words and the way in which they are combined. (Partee, 1995)

"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)

More lenient ⟷ More stringent

```
on(the cat, the mat)

on(the mat, the cat)
```

What does it mean to be "compositional"?

Candidate Definitions

The meaning of a sentence is a function of the meanings of the words and the way in which they are combined. (Partee, 1995)

"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)

More lenient ⟷ More stringent

In line with Montague, Chomsky traditions

What does it mean to be "compositional"?

Candidate Definitions

The meaning of a sentence is a function of the meanings of the words and the way in which they are combined. (Partee, 1995)

"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)

More lenient ⟷ More stringent

In line with Montague, Chomsky traditions

Not <u>obviously</u> compatible with neural networks

What does it mean to be “compositional”?

This Talk

- I will focus on the latter definition: i.e., “compositional” in the stronger, quasi-formal-language sense
- I focus on this definition because:
    - The answer is **non-obvious**, how to go about answering it is non-trivial, and thus it is **interesting**!
    - Some aspects of human cognition likely require some aspects of this type of representation (e.g., we can do math, and we can write code, so, at least *sometimes,* **we do logic**)
    - AI will be used for many things, **not *just* replicating humans**. It’s relevant whether a computational model can implement such a system, whether or not humans do it.

What does it mean to be "compositional"?

Disclaimers on my personal opinions

- I do <u>**not**</u> use this definition because.
  - I believe these representations are **"right"** and others are "wrong".
  - I believe that these representations are necessarily ***required*** for "human-level" language performance

What does it mean to be "compositional"?

Disclaimers on my personal opinions

- I adopt a liberal version of this definition. So, let's concede:
  - Representations can (should!) be **continuous**. This isn't a debate about discrete vs. continuous, its about compositional vs. non-compositional.
  - Syntax-driven semantic composition is an important part of the story, its **not the whole story**. Top-down influence/context-dependence is allowed (necessary!). Idiomatic use and memorization is allowed (necessary!). The point is that a competent AI system has to have the **capacity to represent this type of structure somewhere, somehow**

What does it mean to be “compositional”?

This Talk

- Two questions:

What does it mean to be "compositional"?
This Talk

- Two questions:
    1. Can NNs *learn to implement* a classical cognitive architecture?

What does it mean to be "compositional"?

This Talk

- Two questions:

    1. ~~Can~~ Do NNs *learn to implement* a classical cognitive architecture?

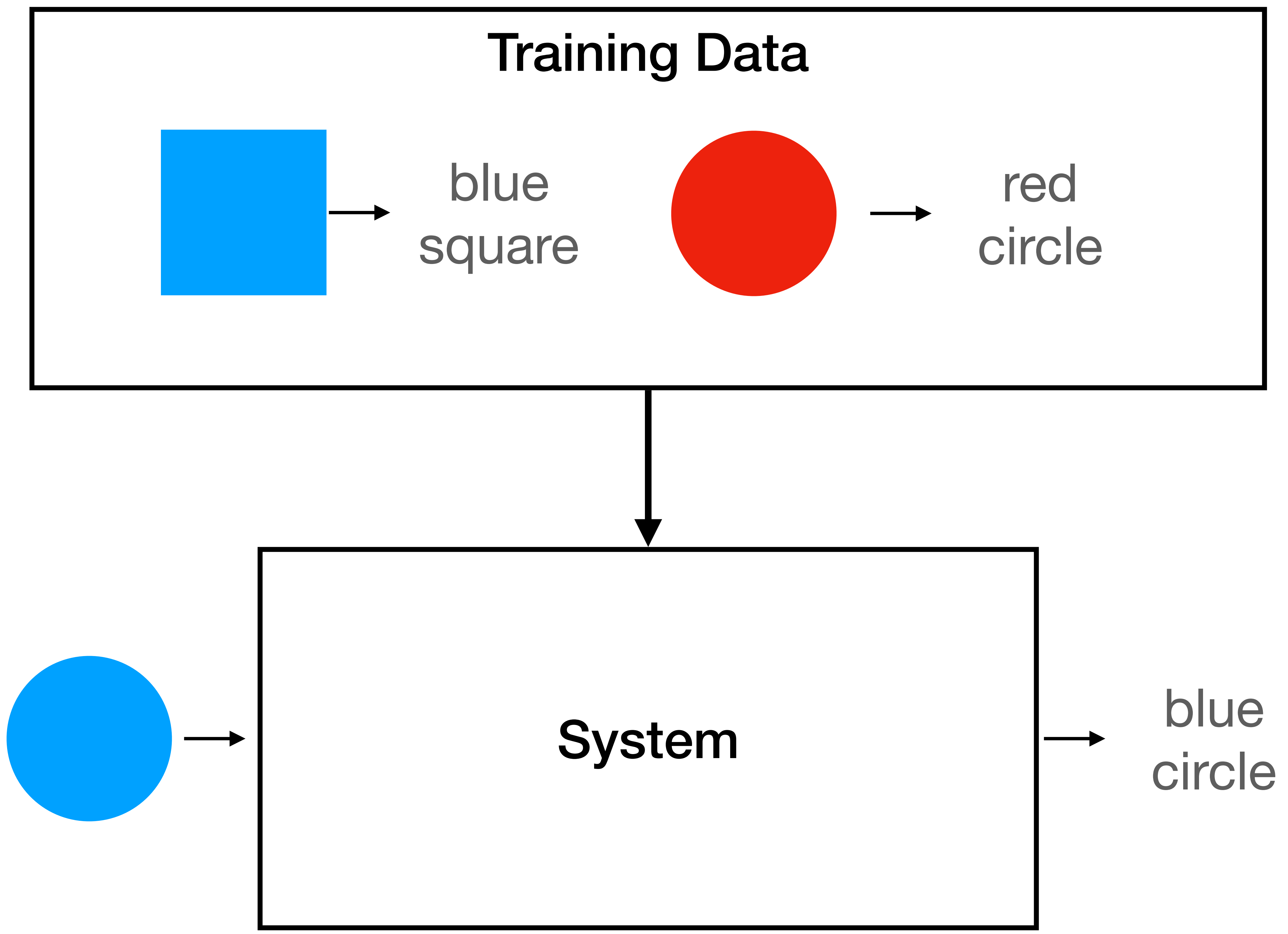
What does it mean to be “compositional”?

This Talk

- Two questions:
    1. ~~Can~~ Do NNs *learn to implement* a classical cognitive architecture?
    2. If so, how would we know?

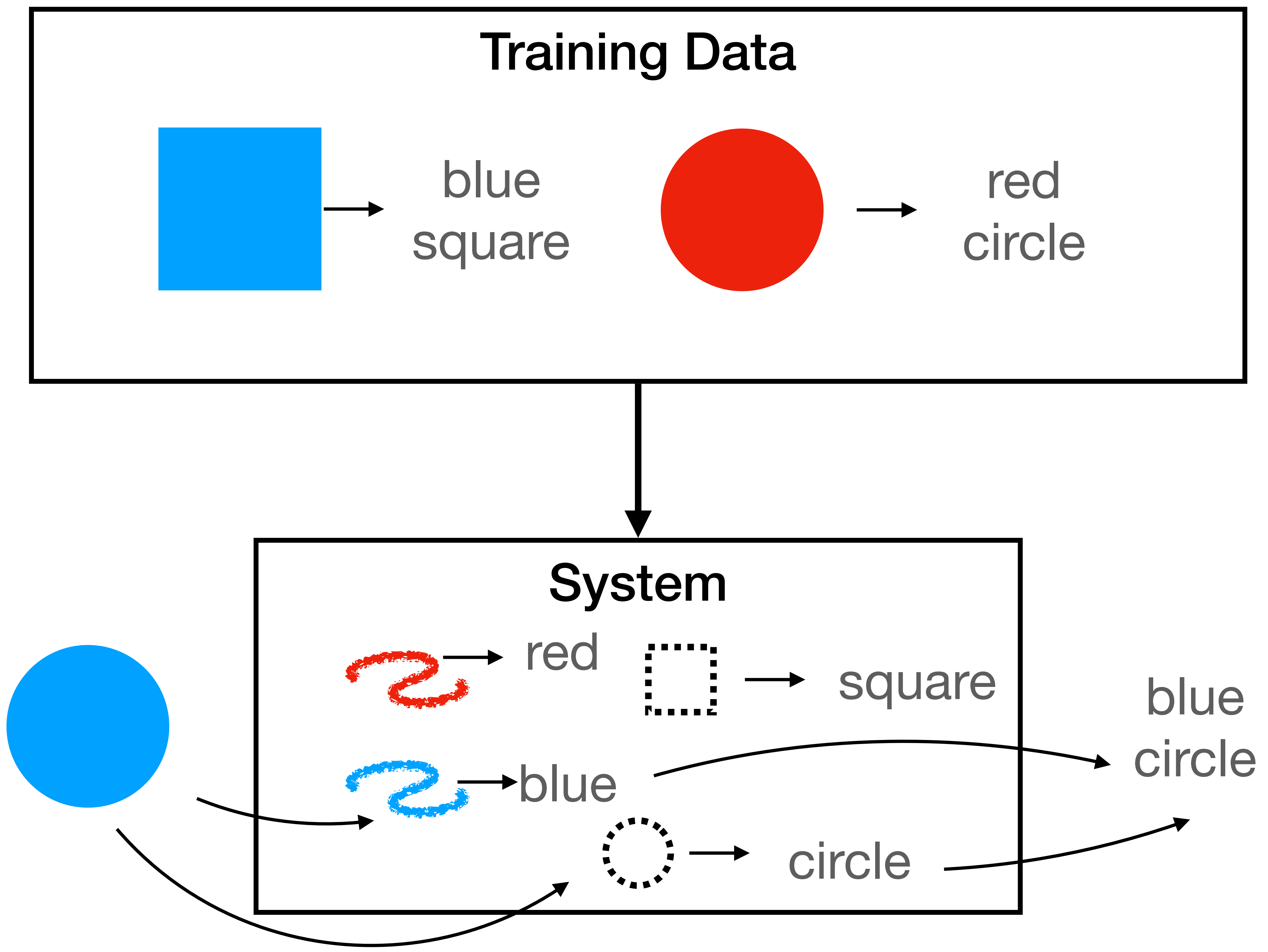
Evaluating compositionality via behavior

Systematic Generalization Tasks



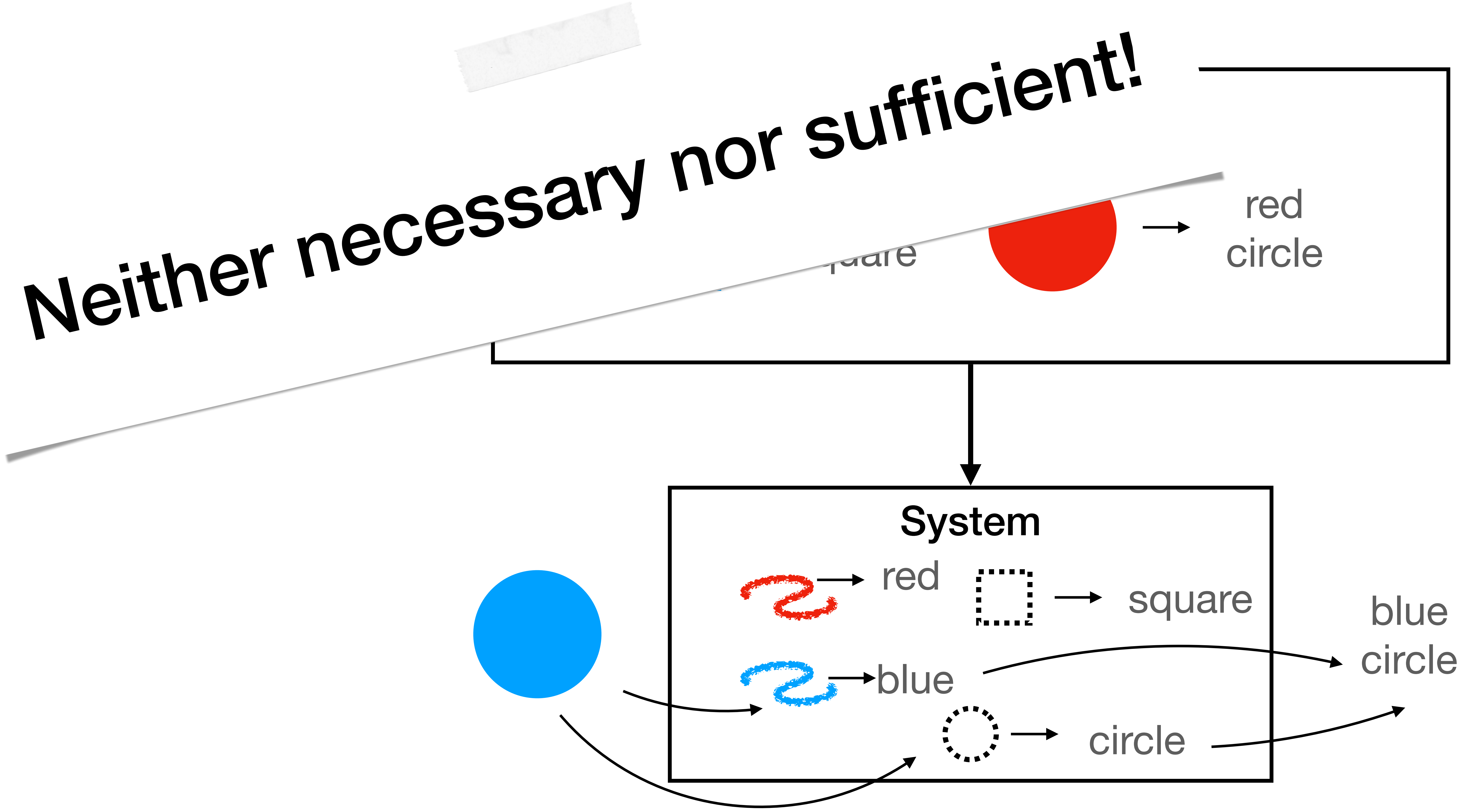
Evaluating compositionality via behavior

Systematic Generalization Tasks



Evaluating compositionality via behavior

Systematic Generalization Tasks

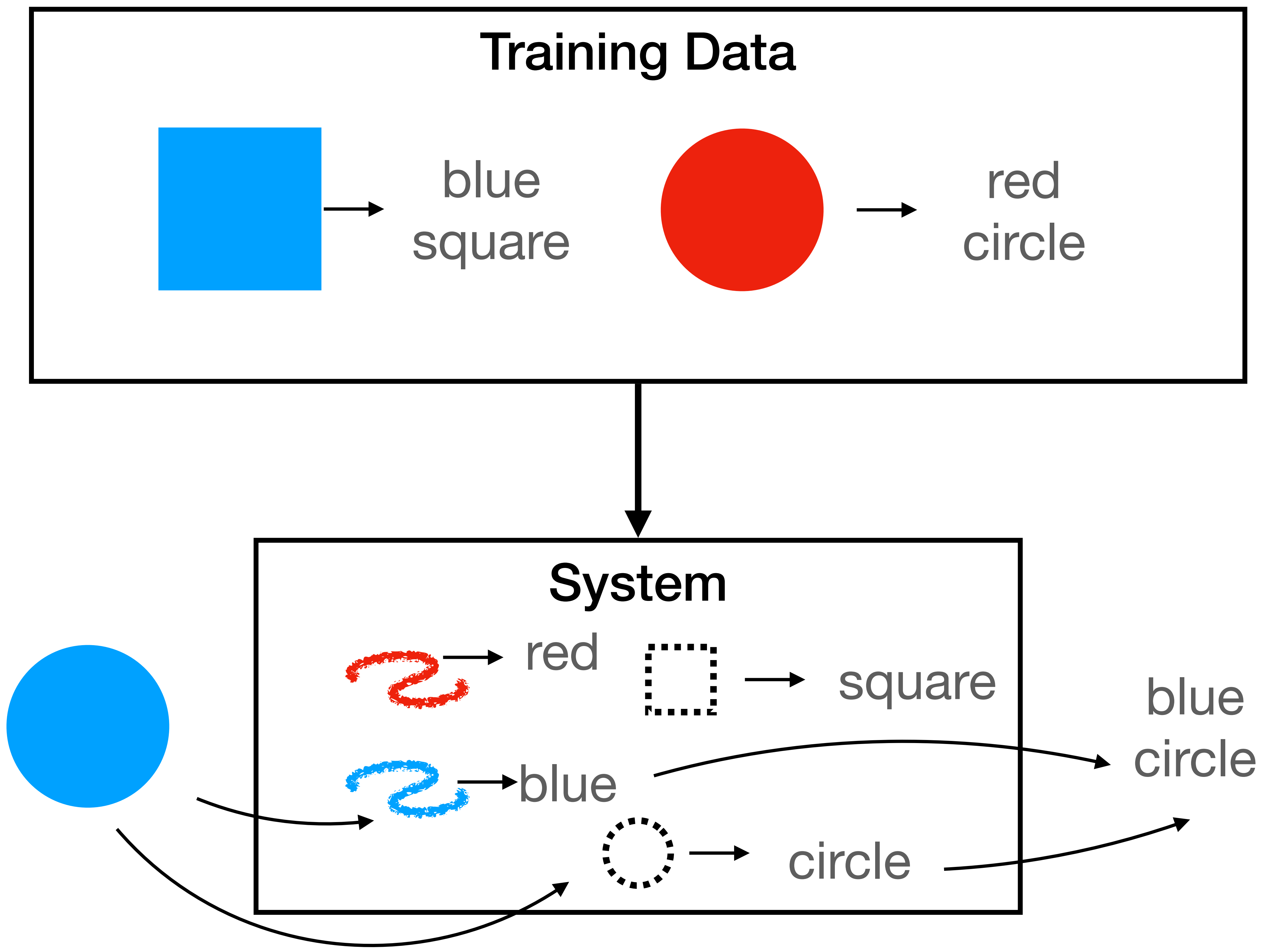


Evaluating compositionality via behavior

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

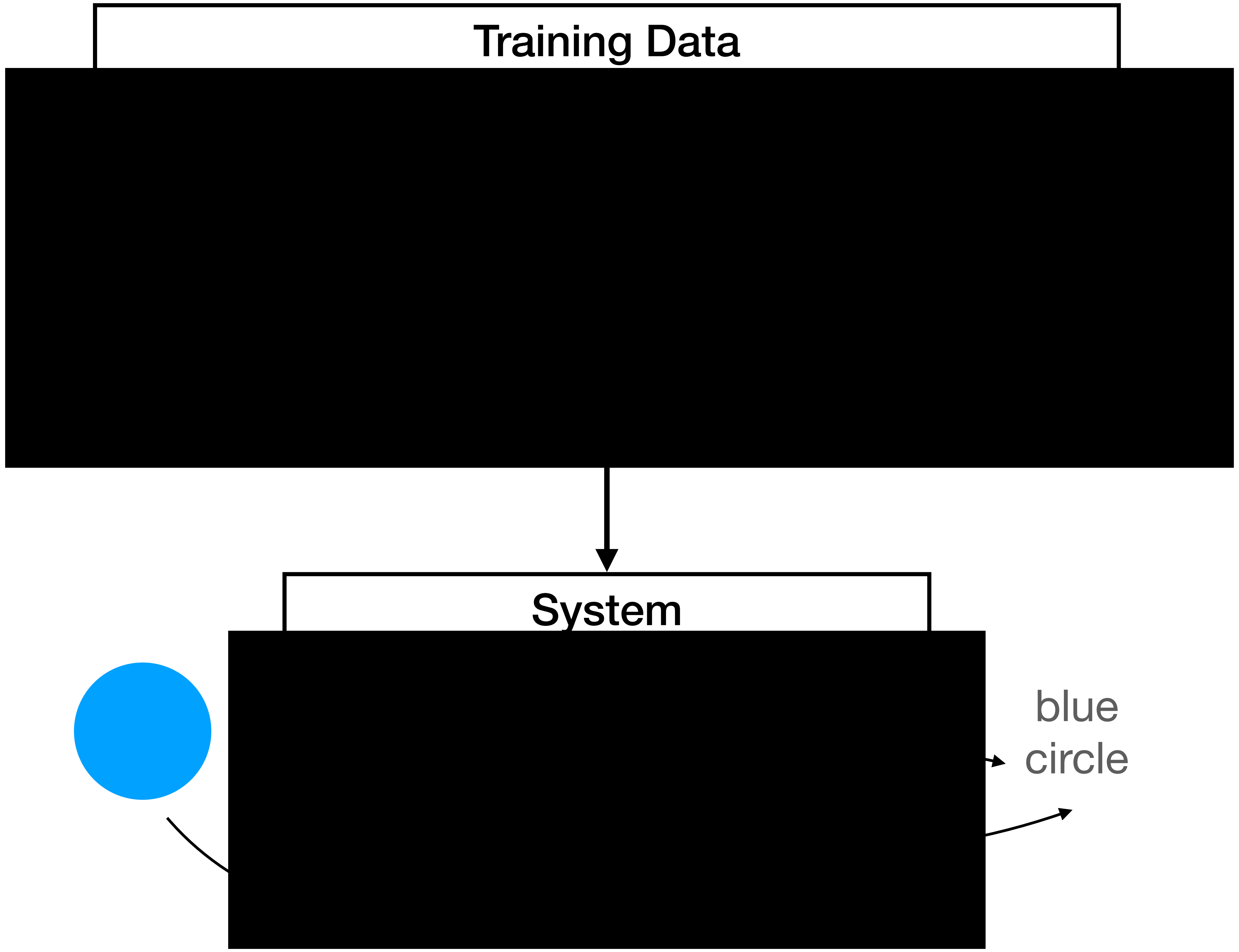
Evaluating compositionality via behavior

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



Evaluating compositionality via behavior

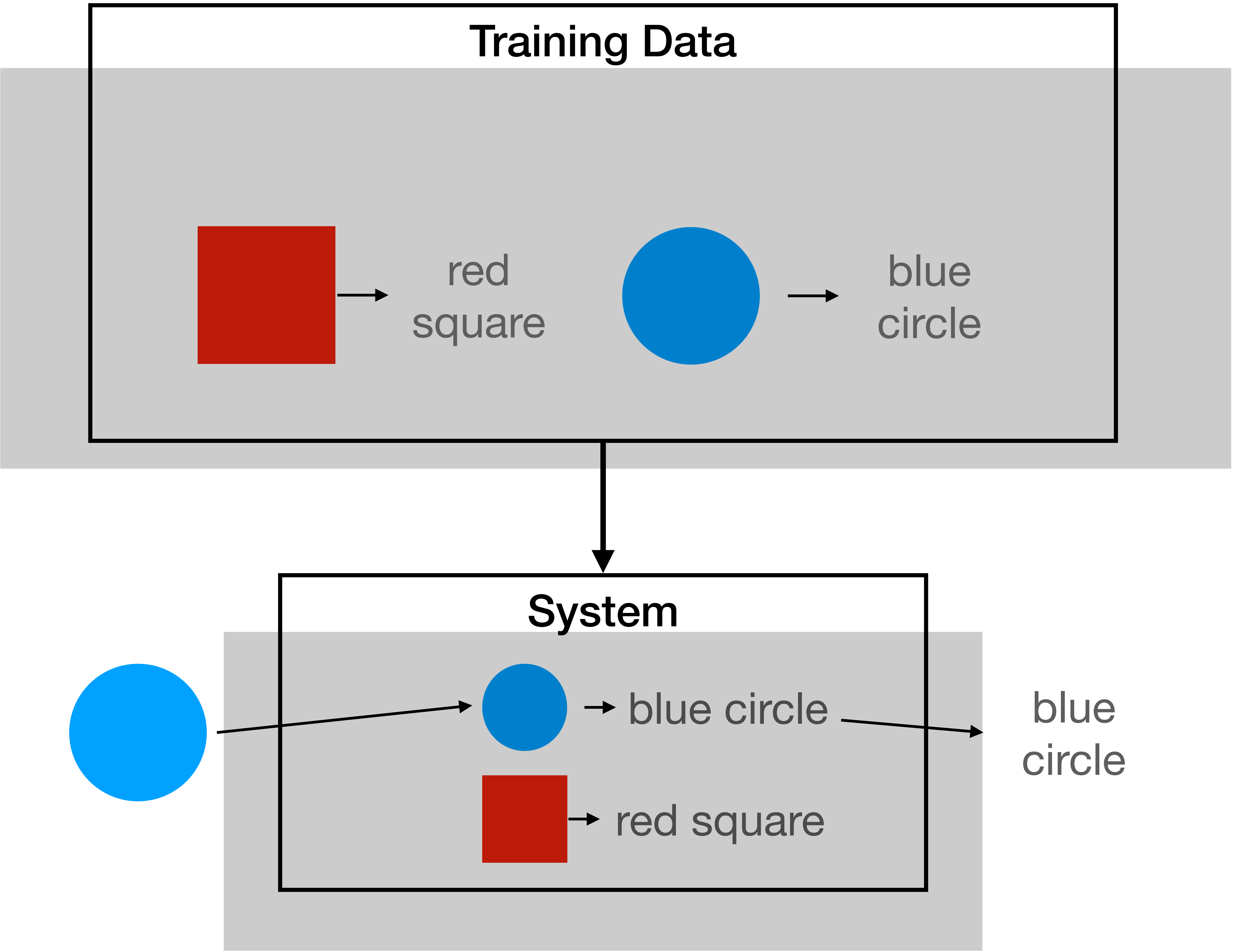
Not Sufficient: Models that don’t meet our definition can still succeed



Issue #1:
For today’s models, we often can’t inspect the training data directly. (Even when its available, its too large to inspect fully and exactly.)

Evaluating compositionality via behavior

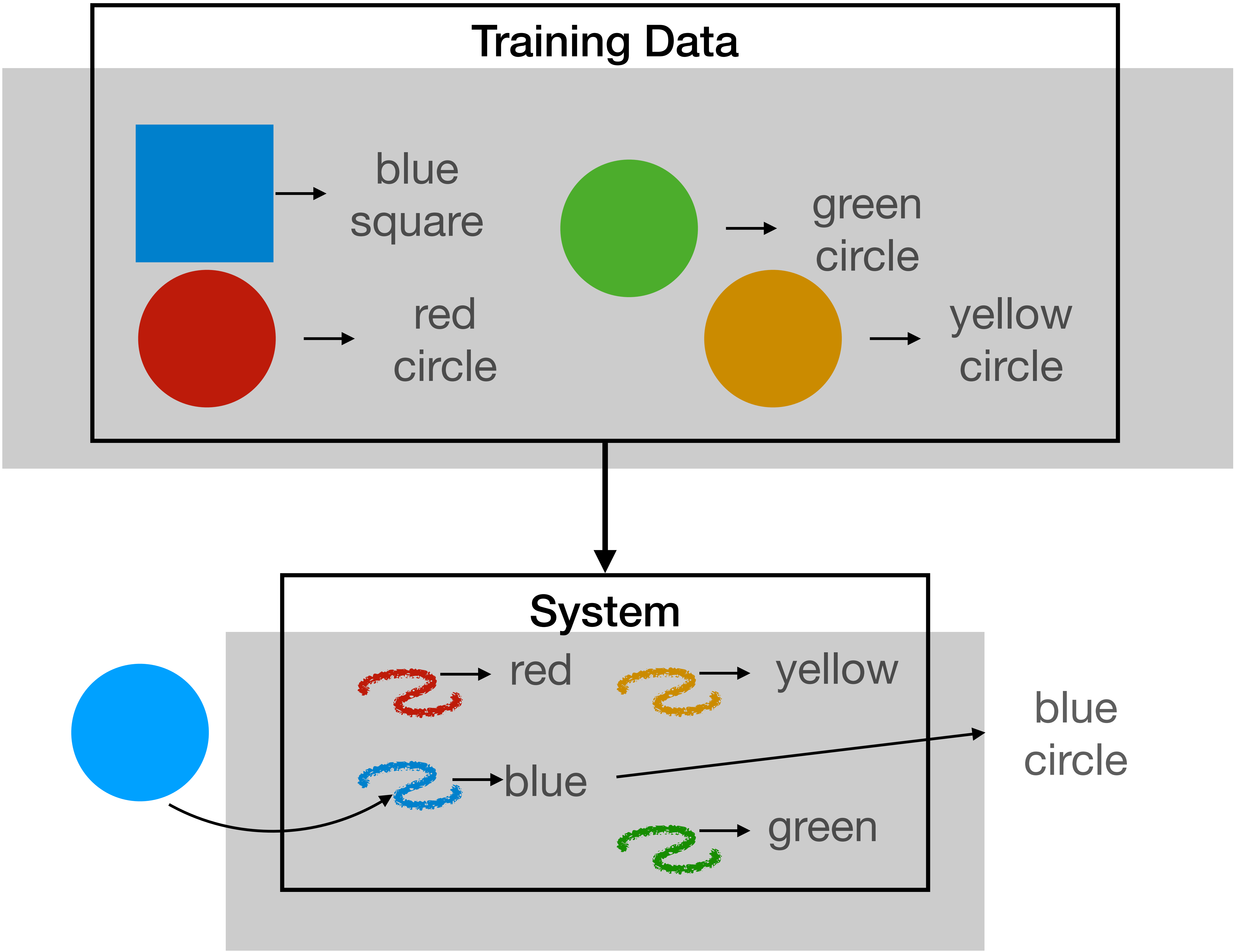
Not Sufficient: Models that don't meet our definition can still succeed



Issue #1:
For today's models, we often can't inspect the training data directly. (Even when its available, its too large to inspect fully and exactly.)

Evaluating compositionality via behavior

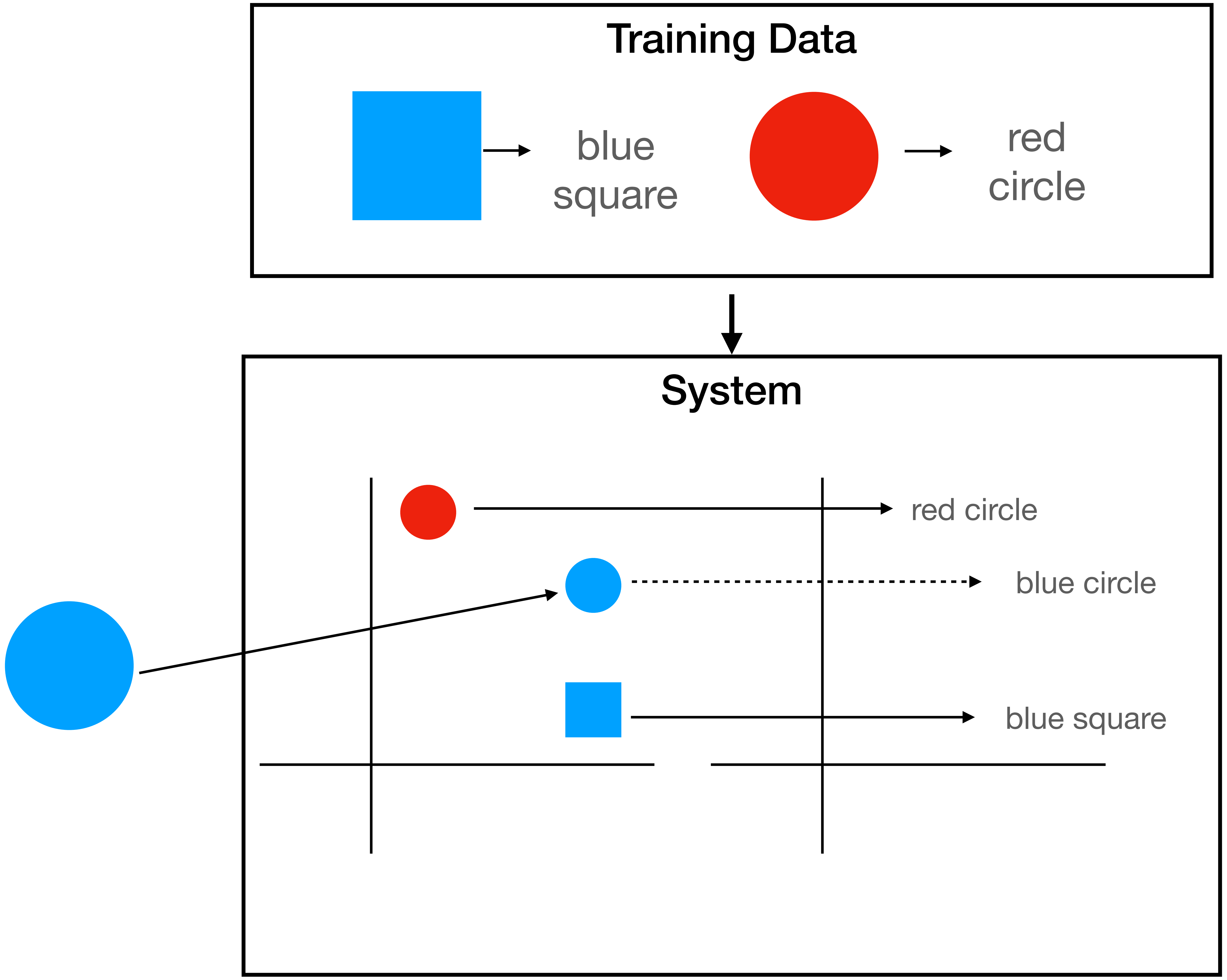
Not Sufficient: Models that don't meet our definition can still succeed



Issue #1:
For today's models, we often can't inspect the training data directly. (Even when its available, its too large to inspect fully and exactly.)

Evaluating compositionality via behavior

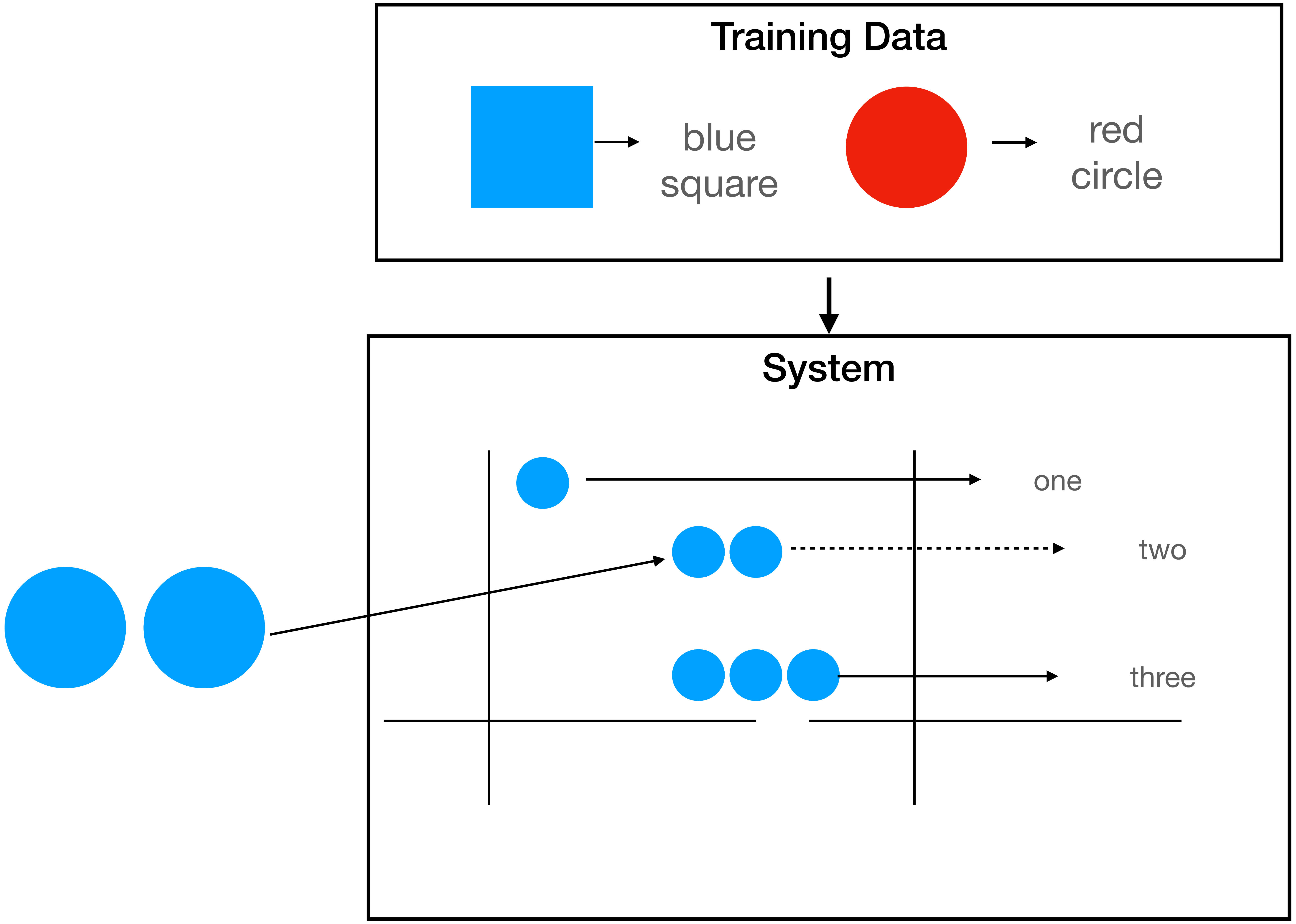
Not Sufficient: Models that don't meet our definition can still succeed



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

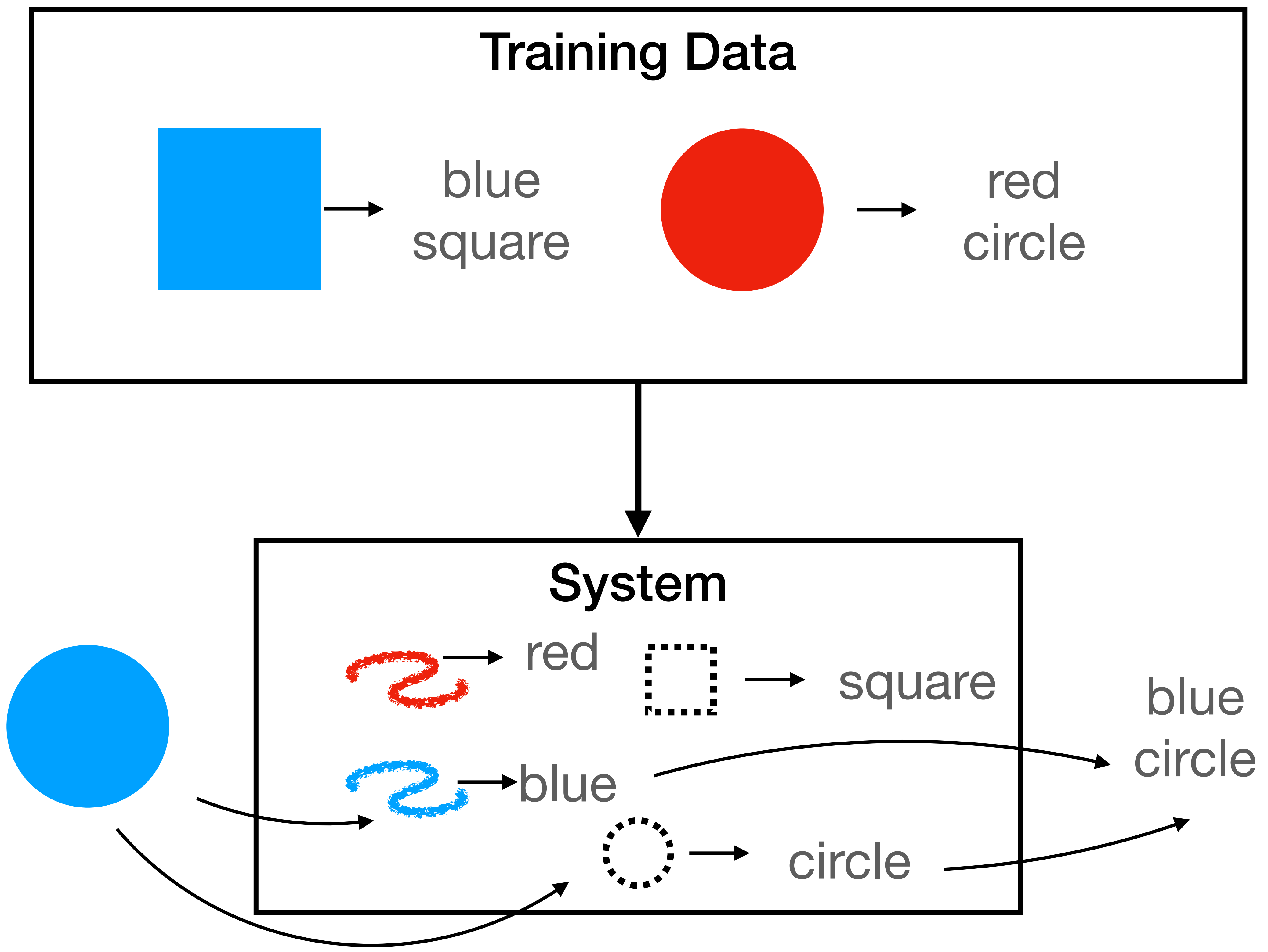
"In between" is not the same as "composed of"

Evaluating compositionality via behavior

Not Necessary: Models that meet our definition could still fail

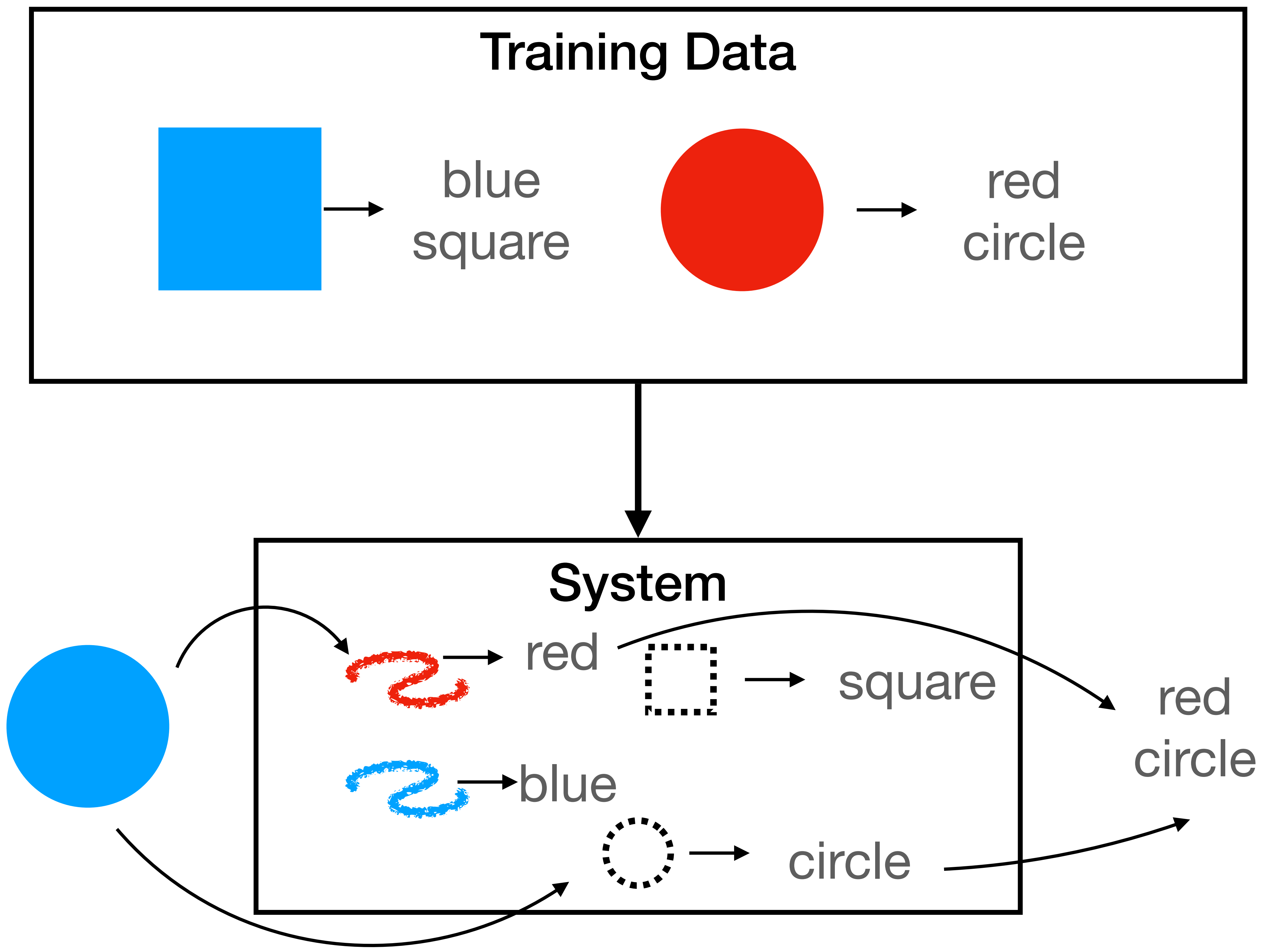
Evaluating compositionality via behavior

Not Necessary: Models that meet our definition could still fail



Evaluating compositionality via behavior

Not Necessary: Models that meet our definition could still fail

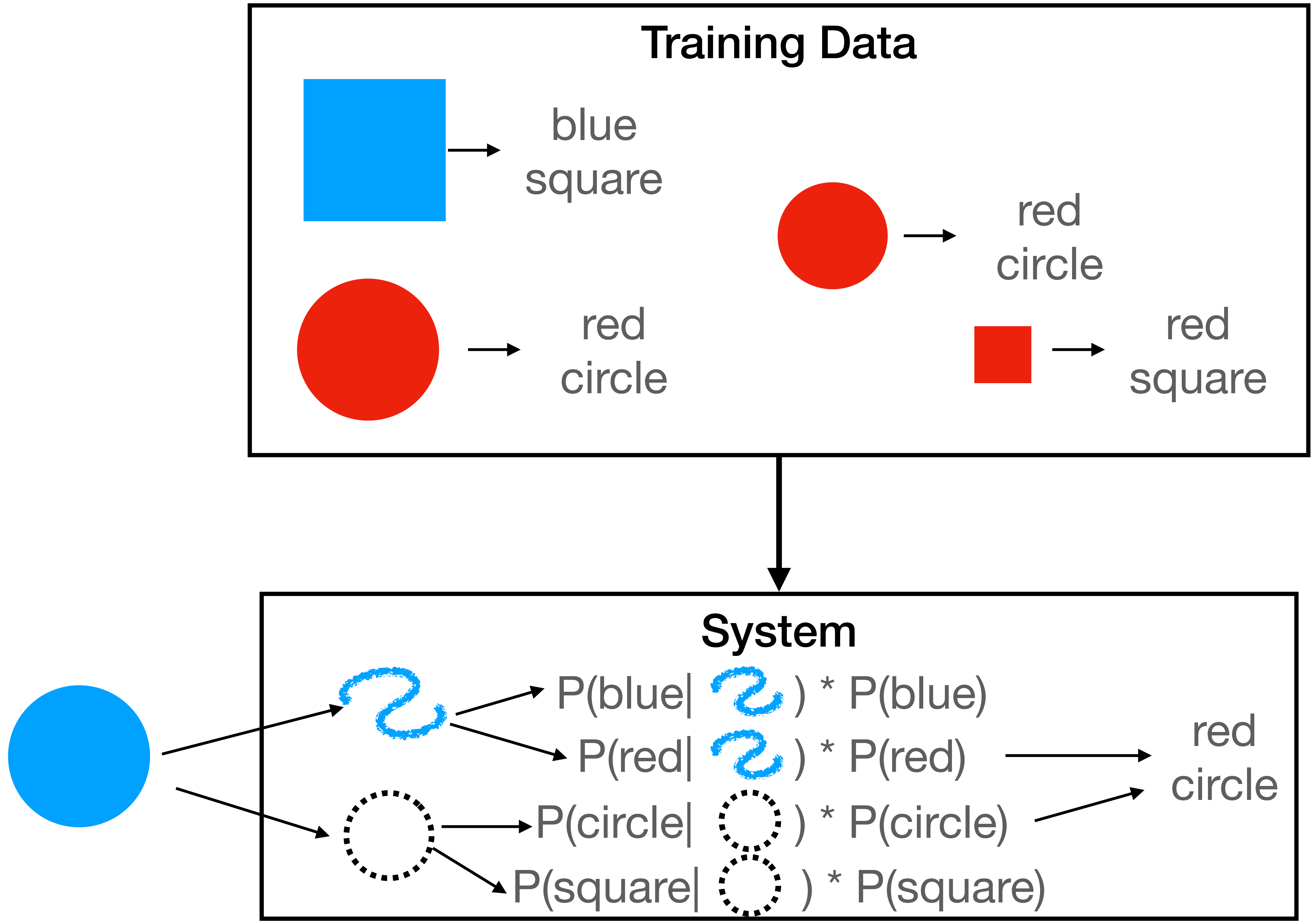


Issue #1: Compositional systems are allowed to make mistakes!

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.

Representation vs. Behavioral Evaluation

- Compositionality (as defined) is a property of **representations**, not behavior
- That doesn't mean behavioral evaluations are not valuable! We of course need to know what models actually do!
- But behavioral evaluations, no matter how carefully constructed, are not **diagnostic** of representations. They alone can't answer our question.
- We need ways to **directly inspect** the internal representations of the model

Representation vs. Behavioral Evaluation

What is a “representational” evaluation?

- **Empirical measures** defined over something other than model inputs and outputs
- Some are slight extensions of behavioral tests, e.g.,
    - Learning Curves: when is one skill acquired relative to another?
    - Reaction/Processing Times: how much “work” is required to produce an output?
- Some are more qualitative:
    - Visualization: Which representations are most similar to one another?
    - Feature Attribution: Which features does the model attend to most to make this decision?
- Newer methods (still in development) attempt to discover explicit mechanisms in the network:
    - Probing: Which neuron or combination of neurons carries this information?
    - Interventions (Pruning/Freezing/Splicing): Can we find the piece of the network that corresponds to a specific behavior?

Case Study

Evaluating a NN Vision Model

Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

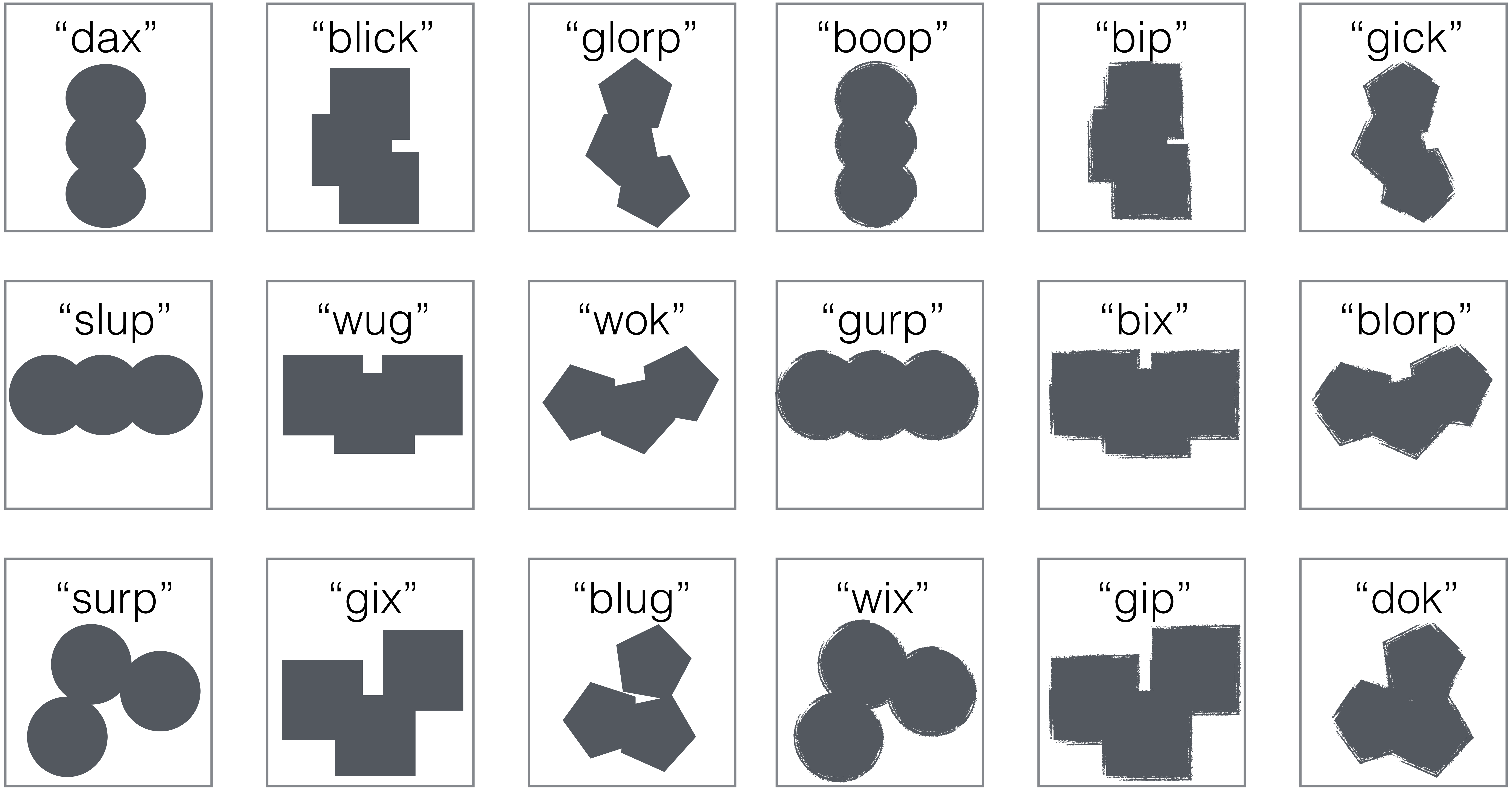
Evaluating a NN Vision Model

Charlie Lovering

Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

Task: Differentiate Simple Visual Concepts

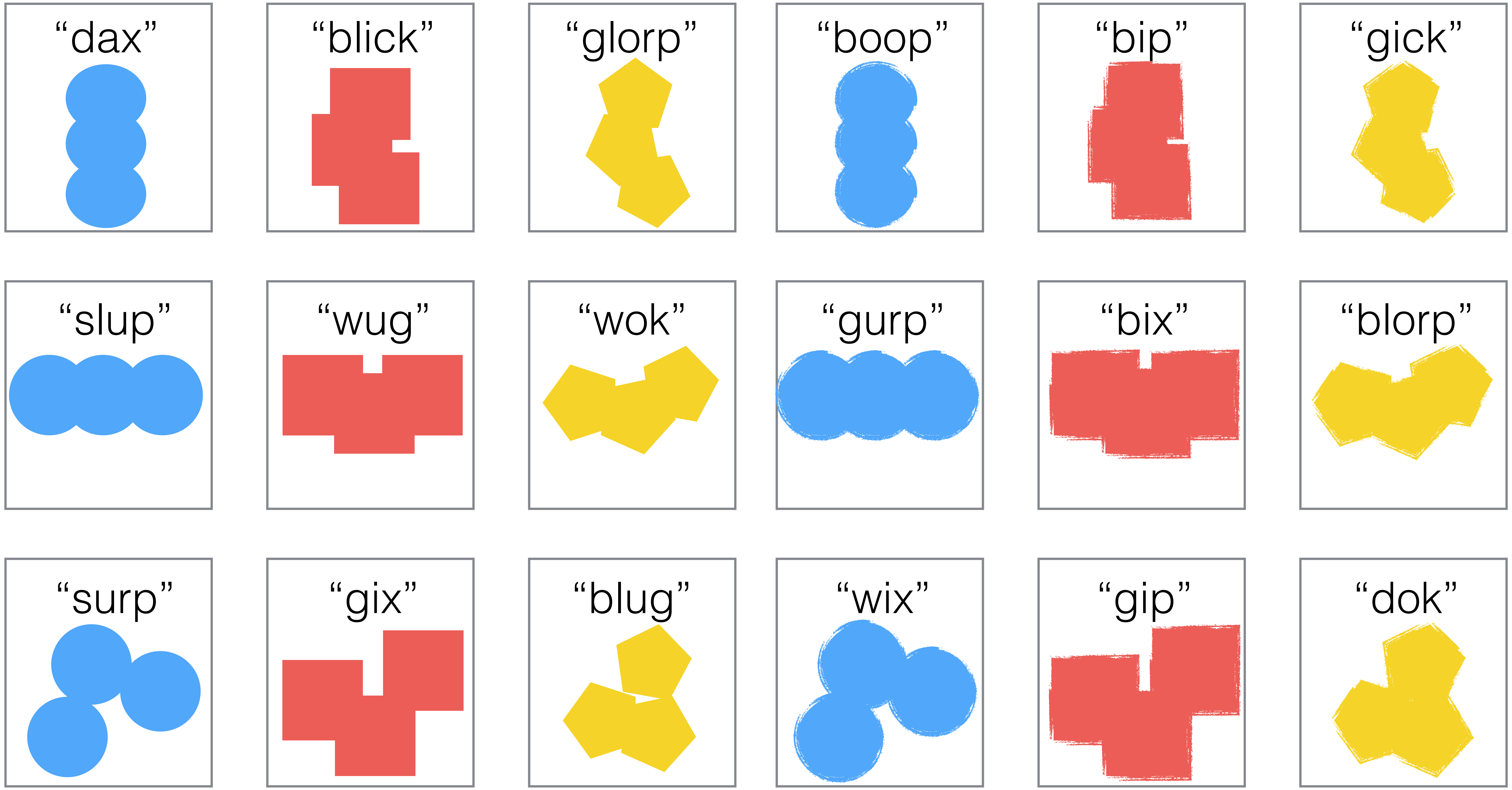


18 high level concepts composed from **8 basic concepts**

Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

Task: Differentiate Simple Visual Concepts

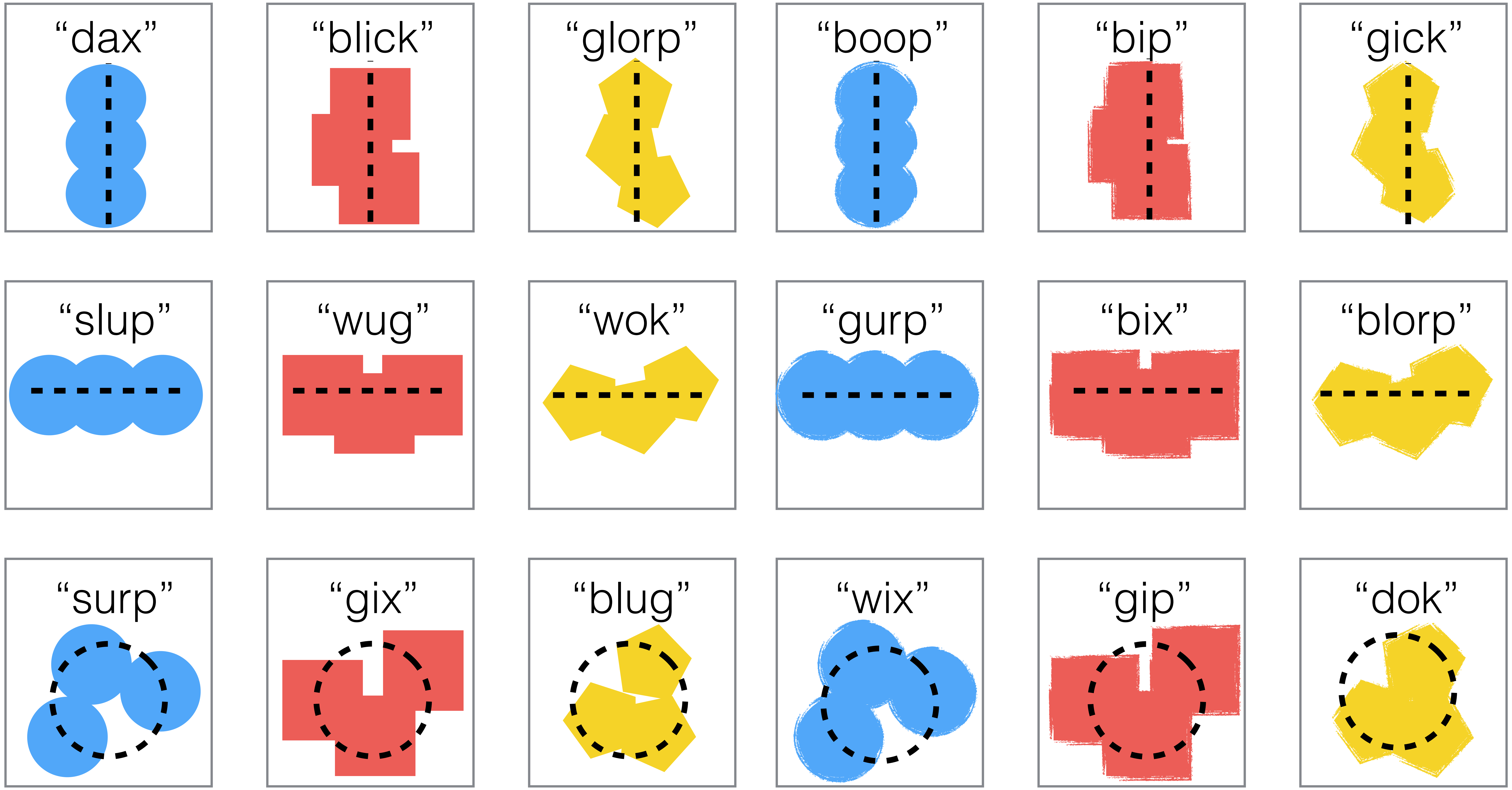


18 high level concepts = {shape}

Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

Task: Differentiate Simple Visual Concepts

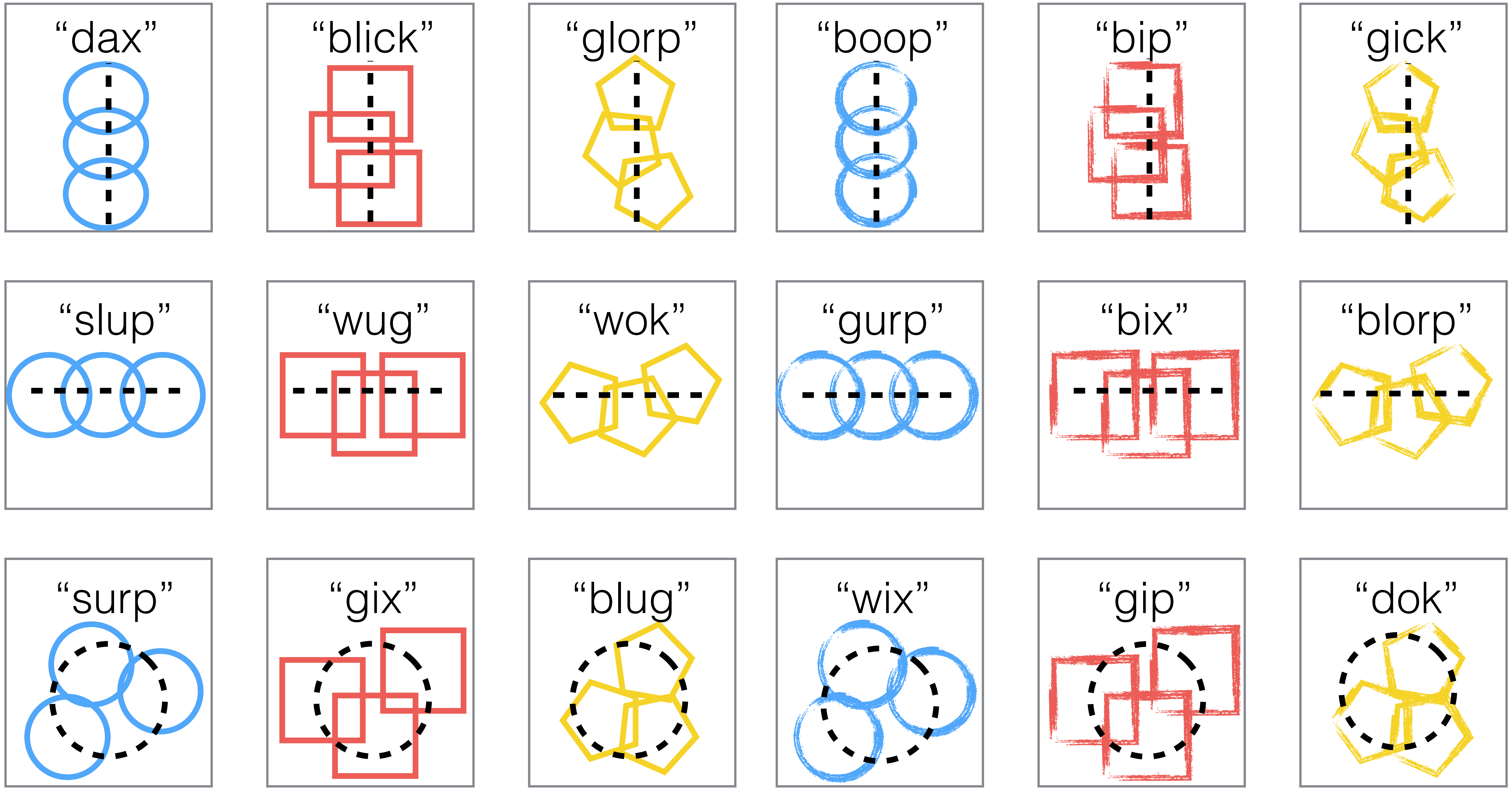


18 high level concepts = {shape} x {layout}

Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

Task: Differentiate Simple Visual Concepts

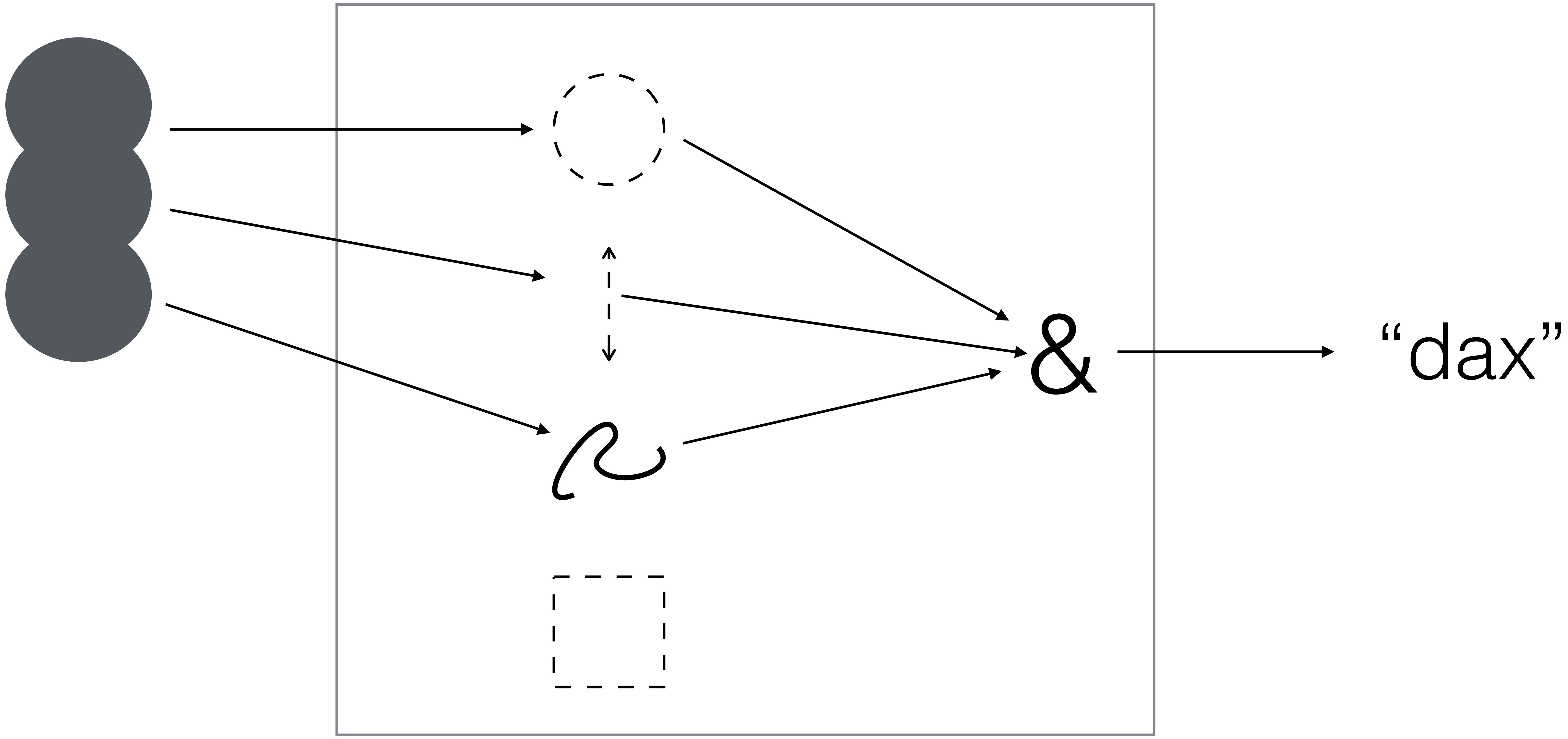


18 high level concepts = {shape} x {layout} x {stroke}

Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

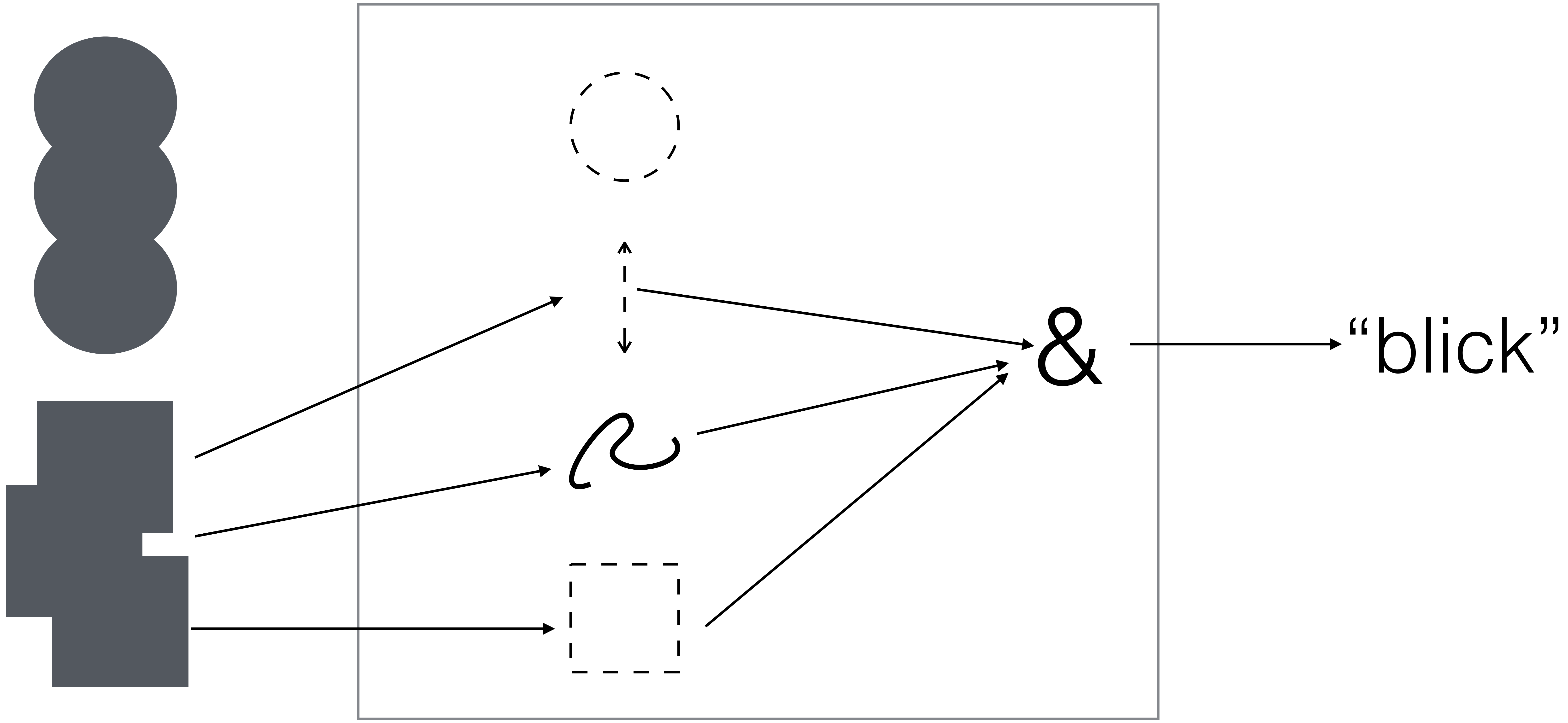
Compositional Conceptual Representation



Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

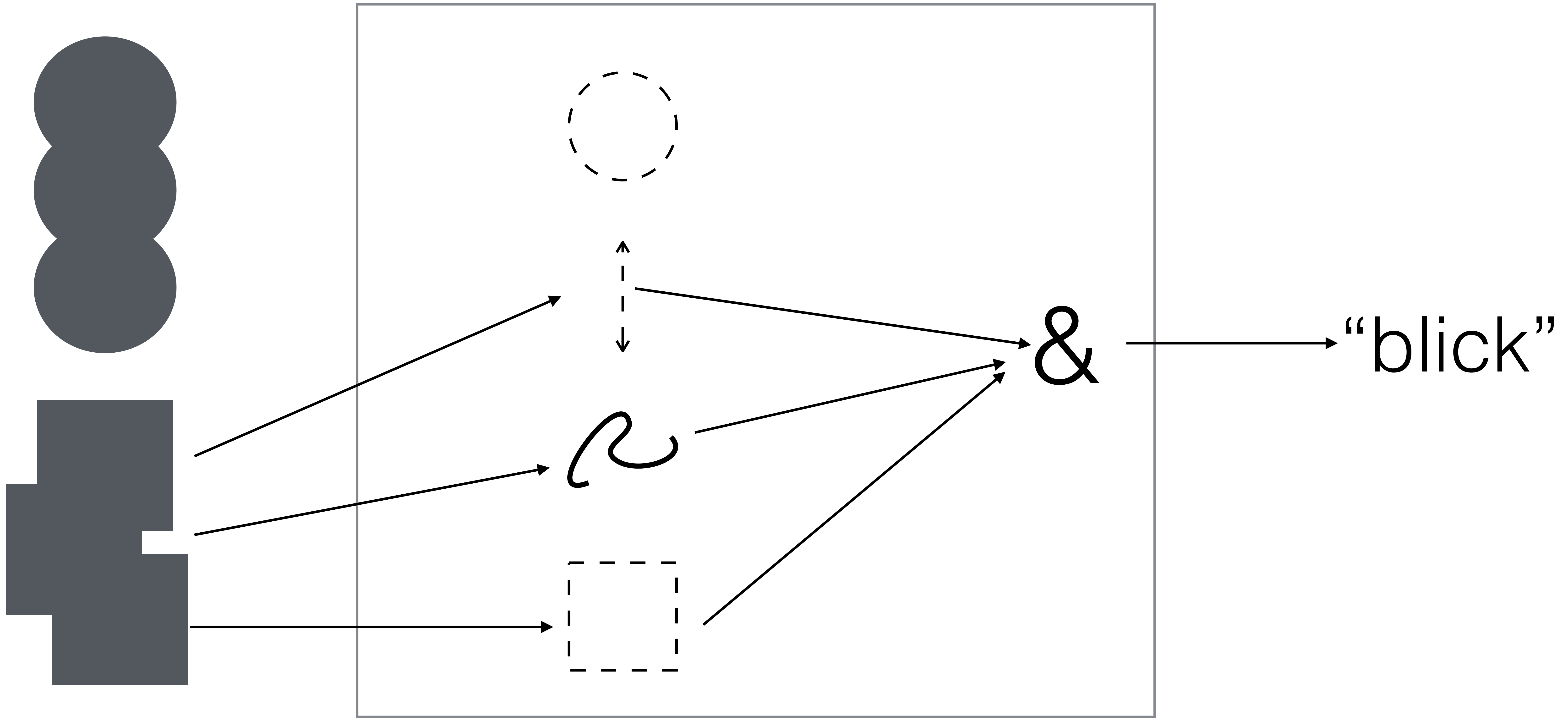
Compositional Conceptual Representation



Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

Compositional Conceptual Representation

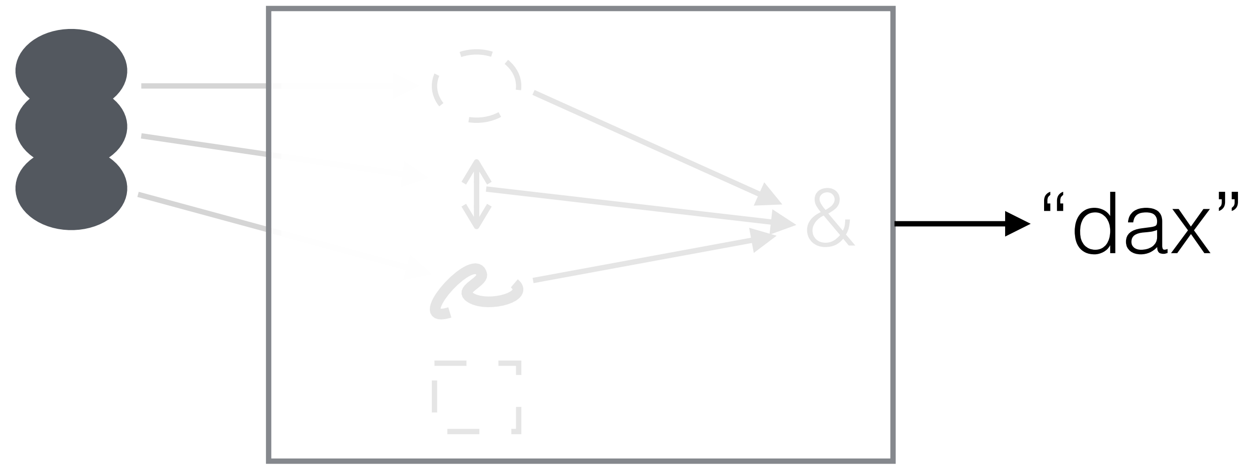
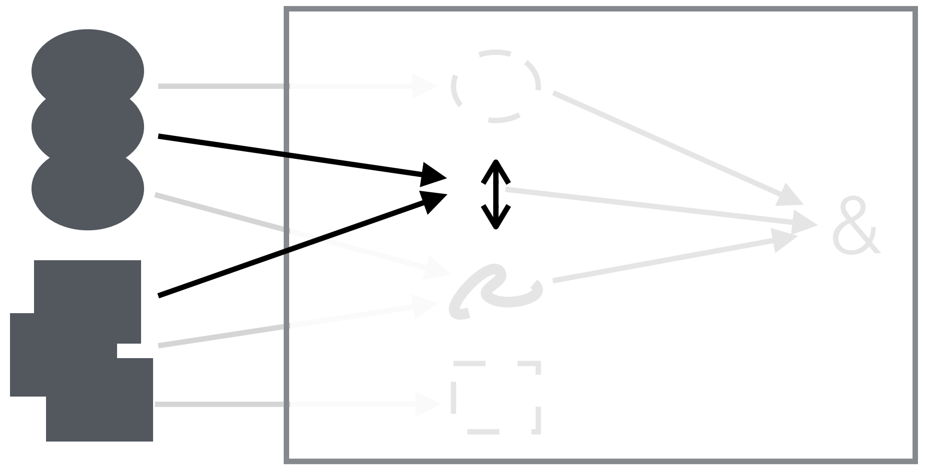
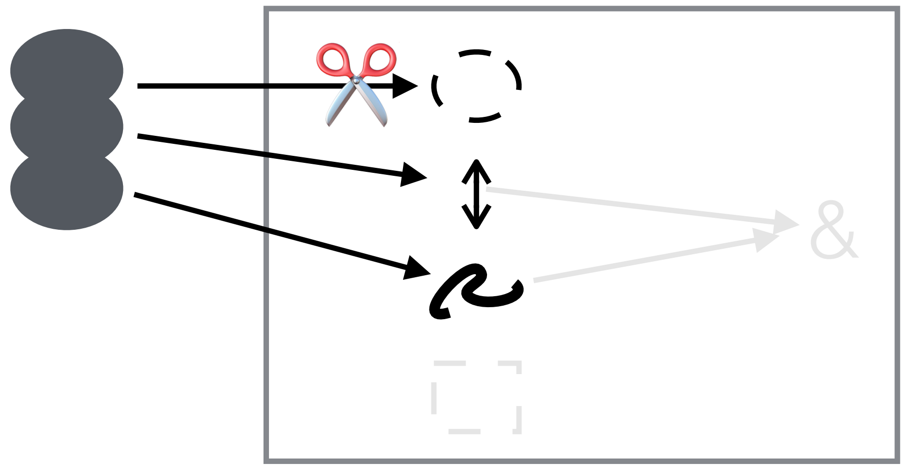
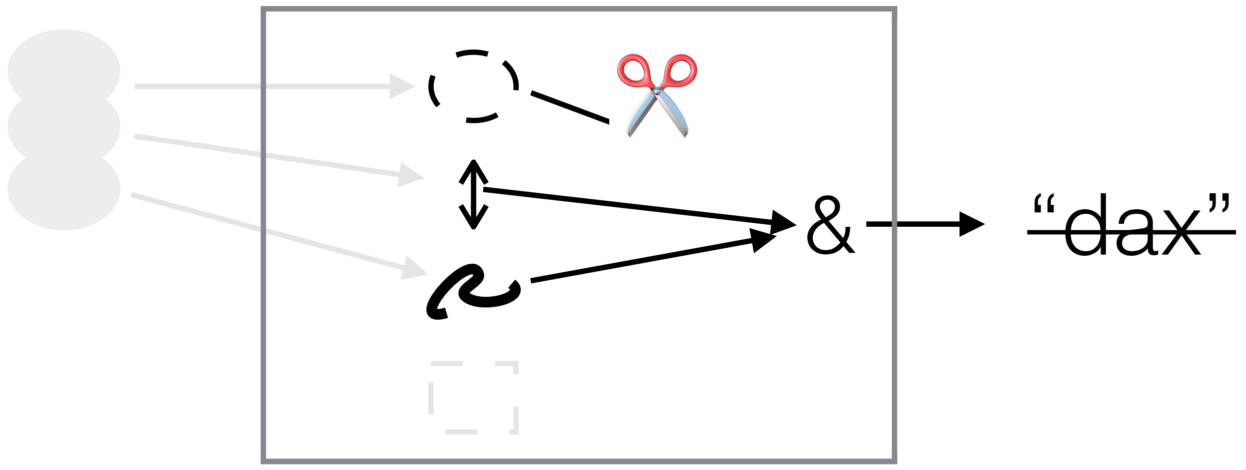


Is our end-to-end NN functionally equivalent to the above system?

Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

Unit Tests

Groundedness	
Reusability of Constituents	
Modularity of Constituents	
Causality of Constituents	

Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

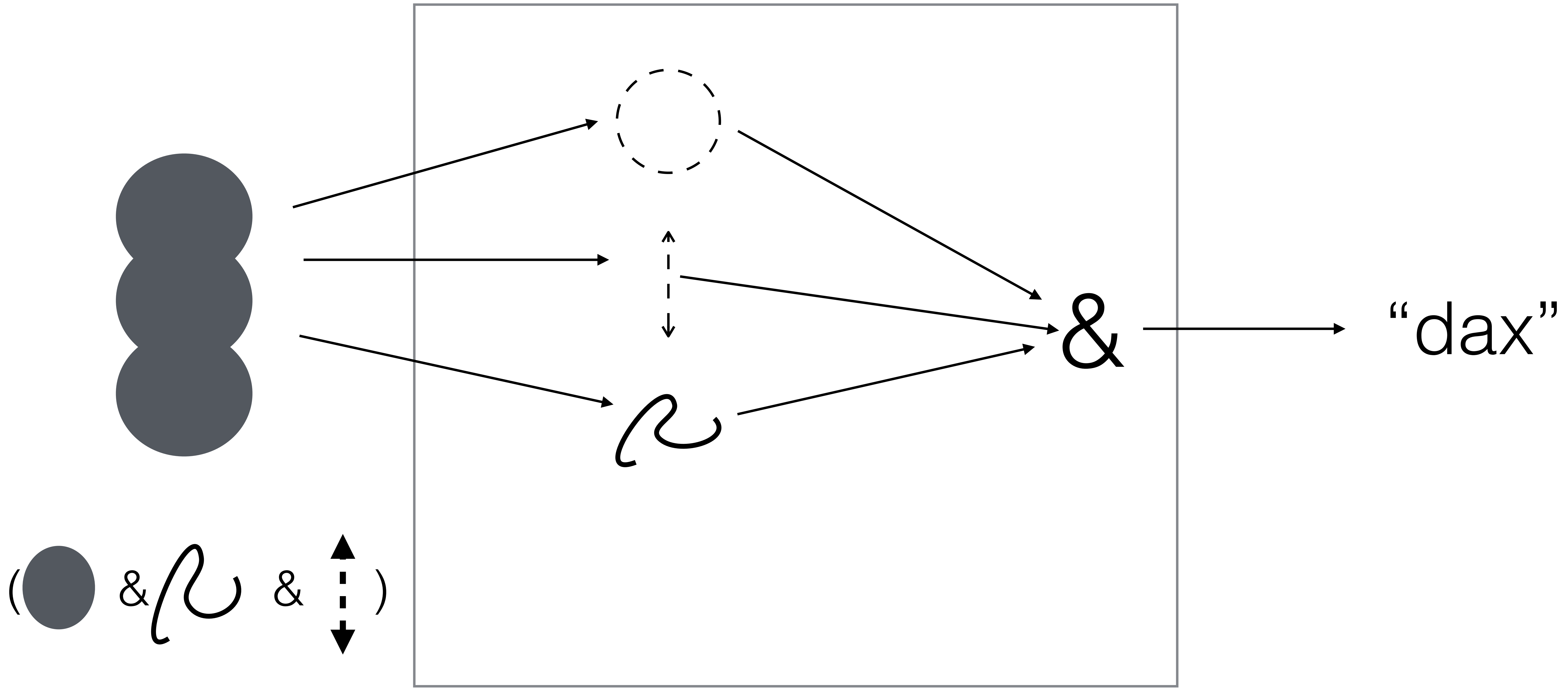
Groundedness

Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

Groundedness

Changes in input lead
to expected changes
in output.

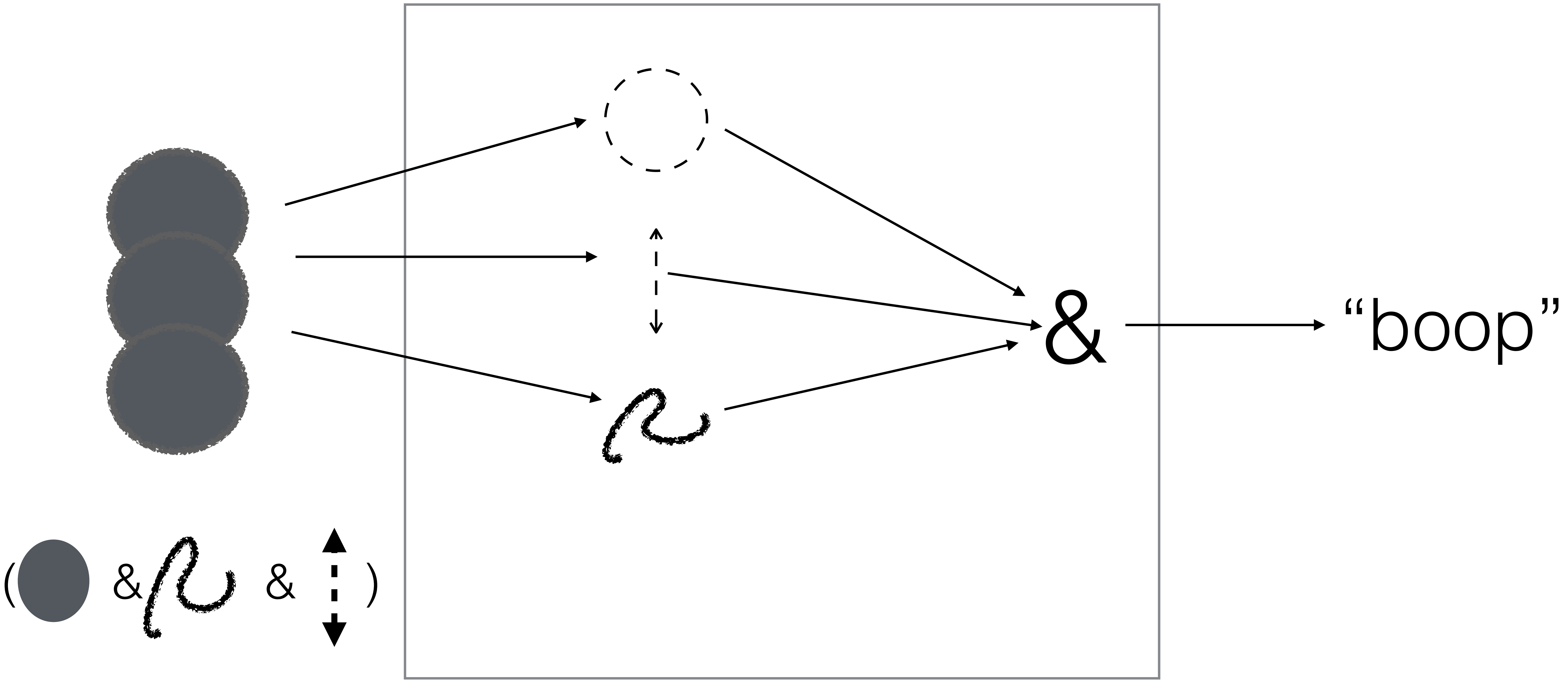


Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

Groundedness

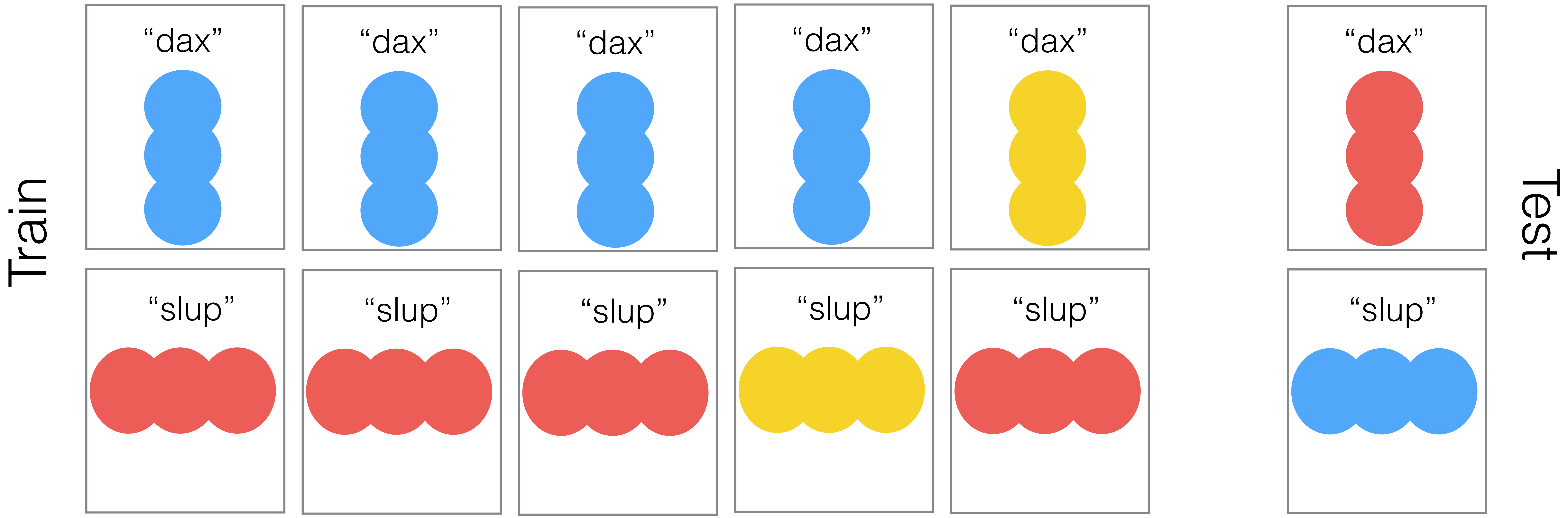
Evaluate using counterfactual minimal pairs



Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

Groundedness: Changes in input -> expected changes in output

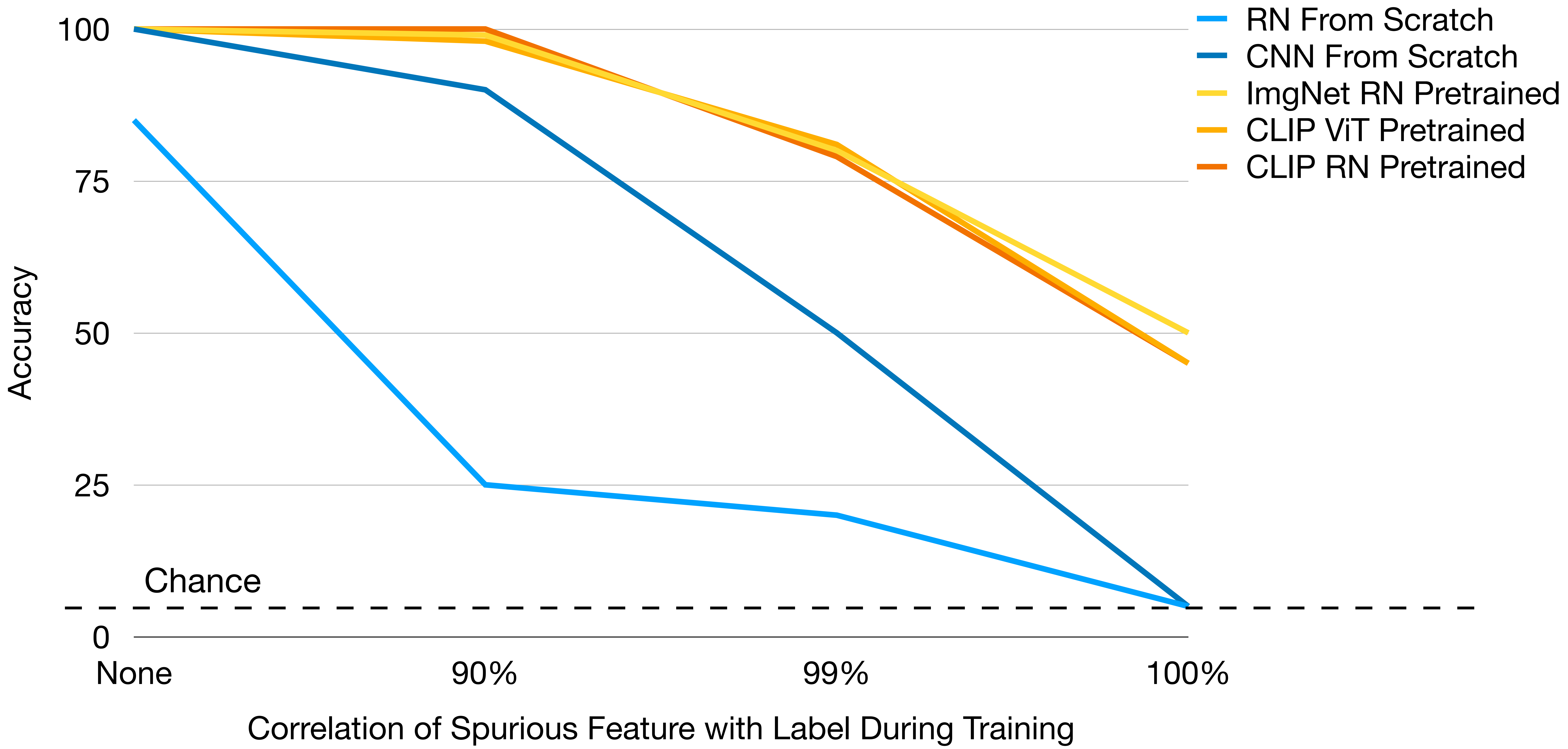


Introduce color as a correlated ("spurious") feature

Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

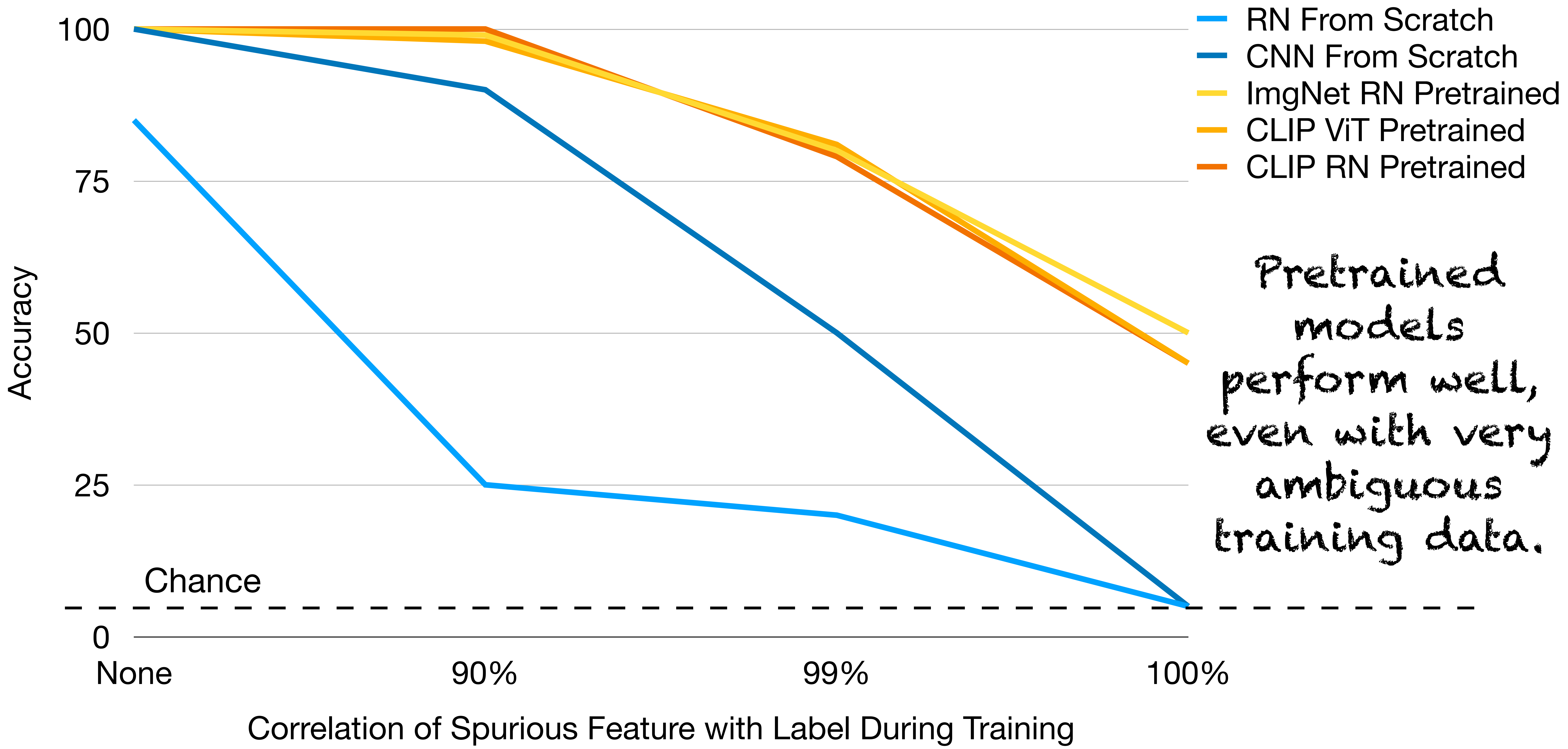
Groundedness: Changes in input -> expected changes in output



Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

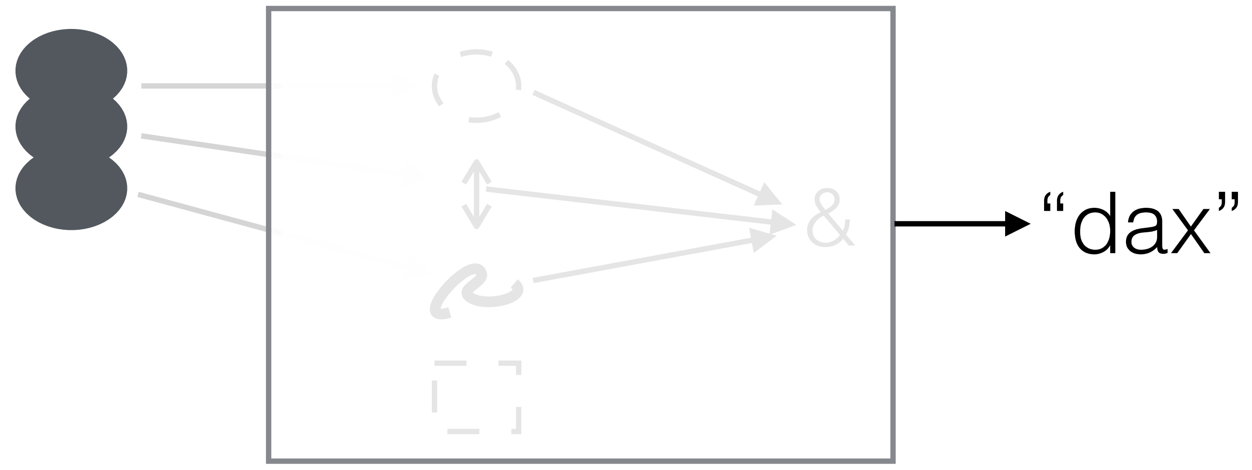
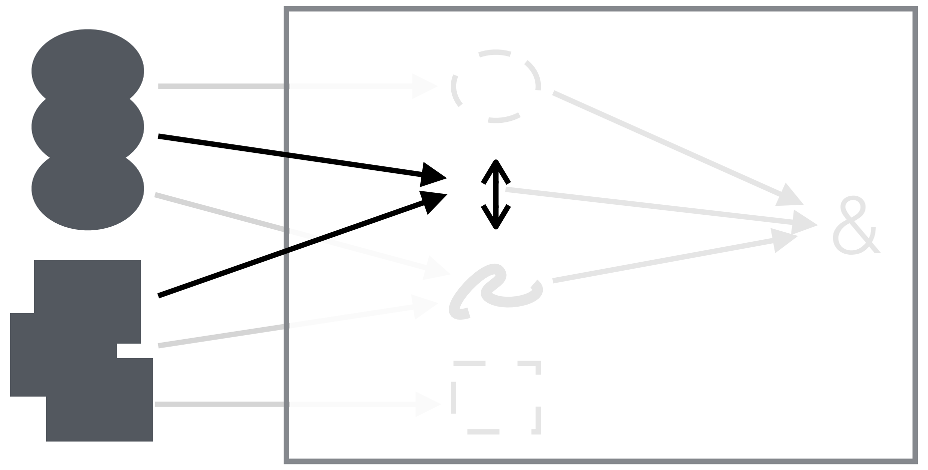
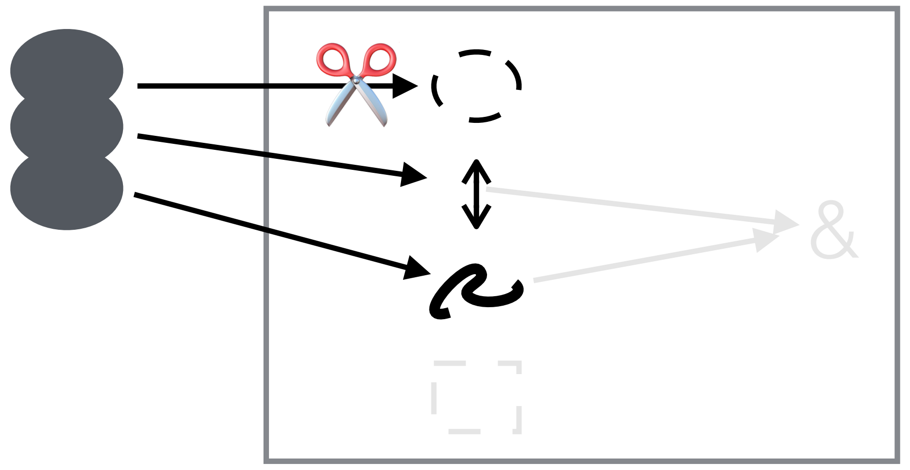
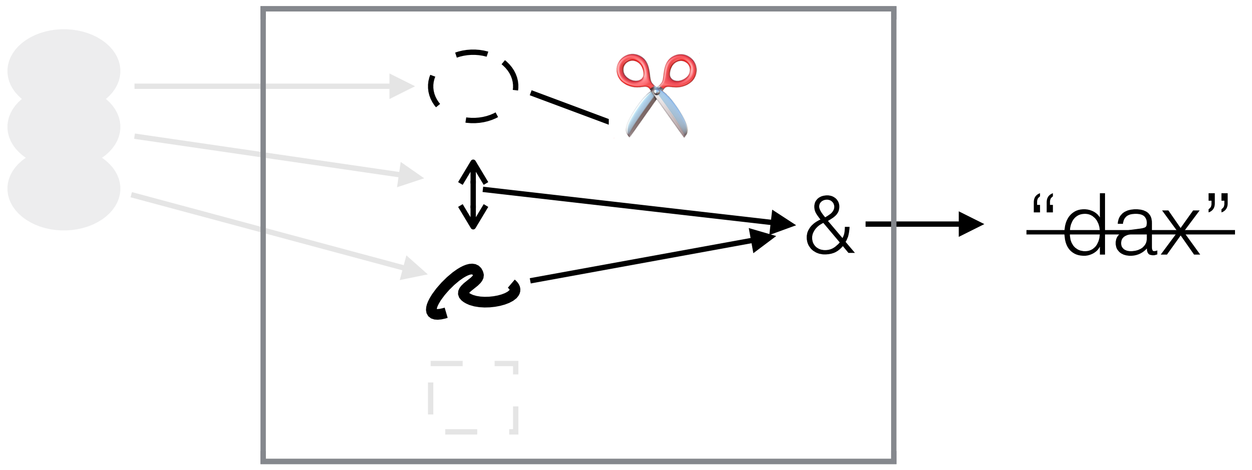
Groundedness: Changes in input -> expected changes in output



Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

Unit Tests

✅	Groundedness	
	Reusability of Constituents	
	Modularity of Constituents	
	Causality of Constituents	

Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

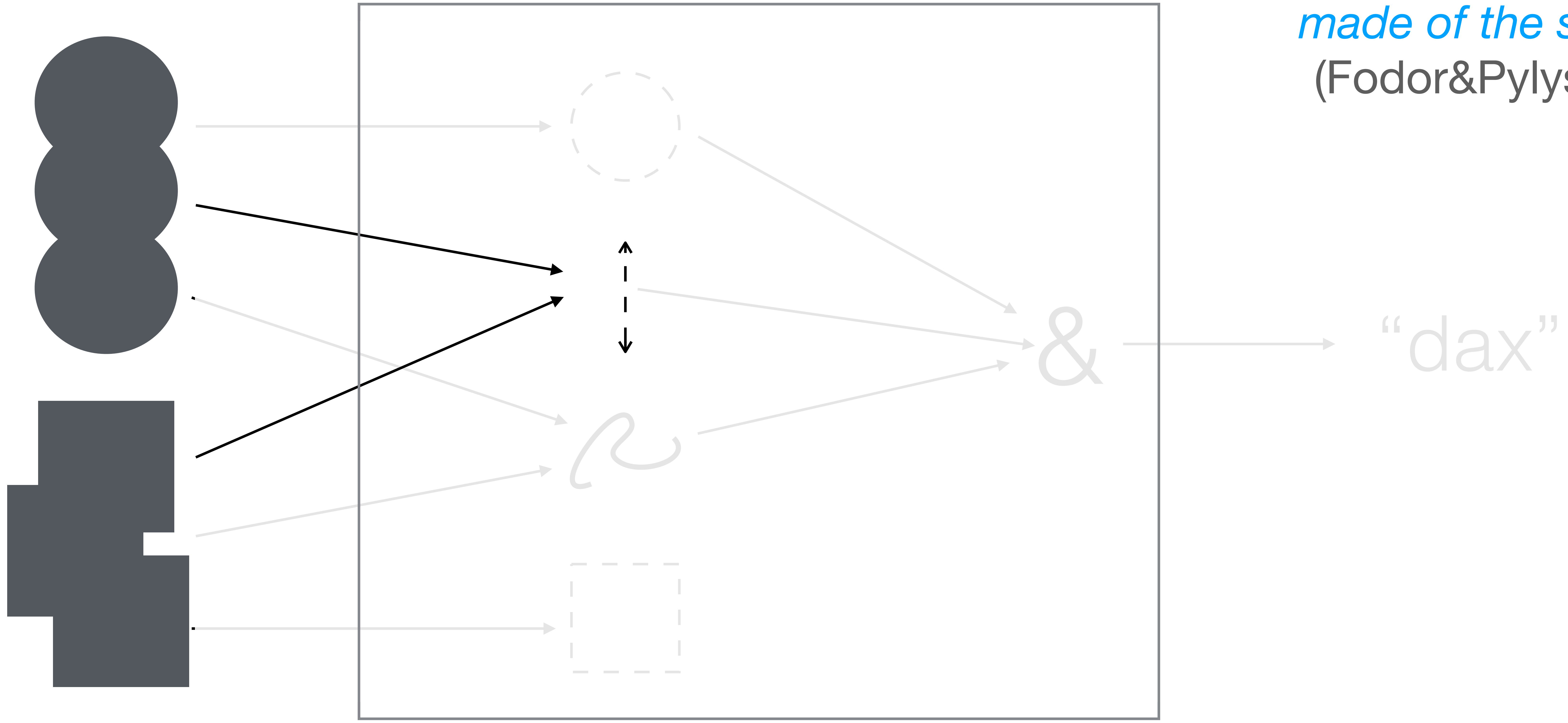
Case Study

Reusability of Constituents

Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

Reusability of Constituents

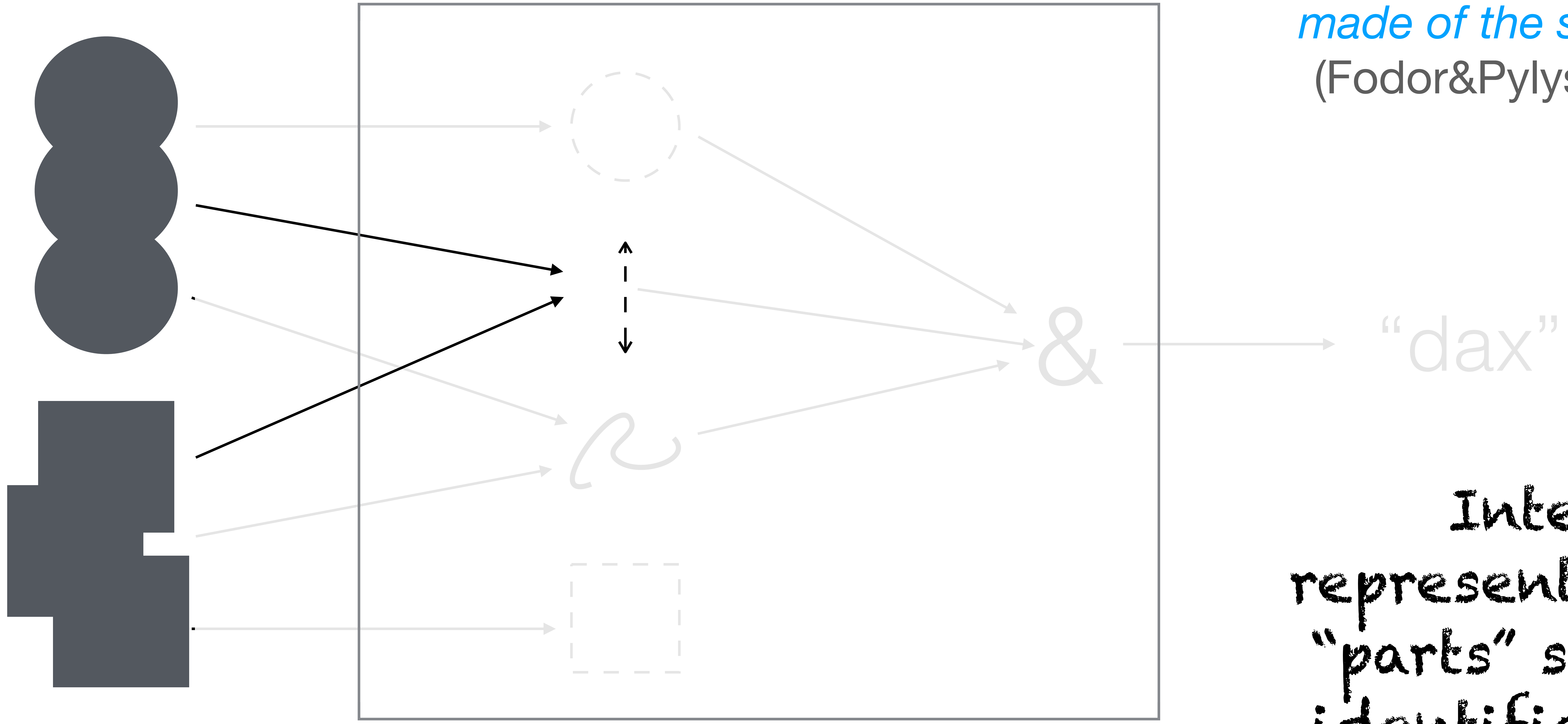


"The ability to produce/understand some sentences is *intrinsically* connected to the ability to produce/understand certain others…[they] *must be made of the same parts*."
(Fodor&Pylyshyn, 1988)

Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

Reusability of Constituents



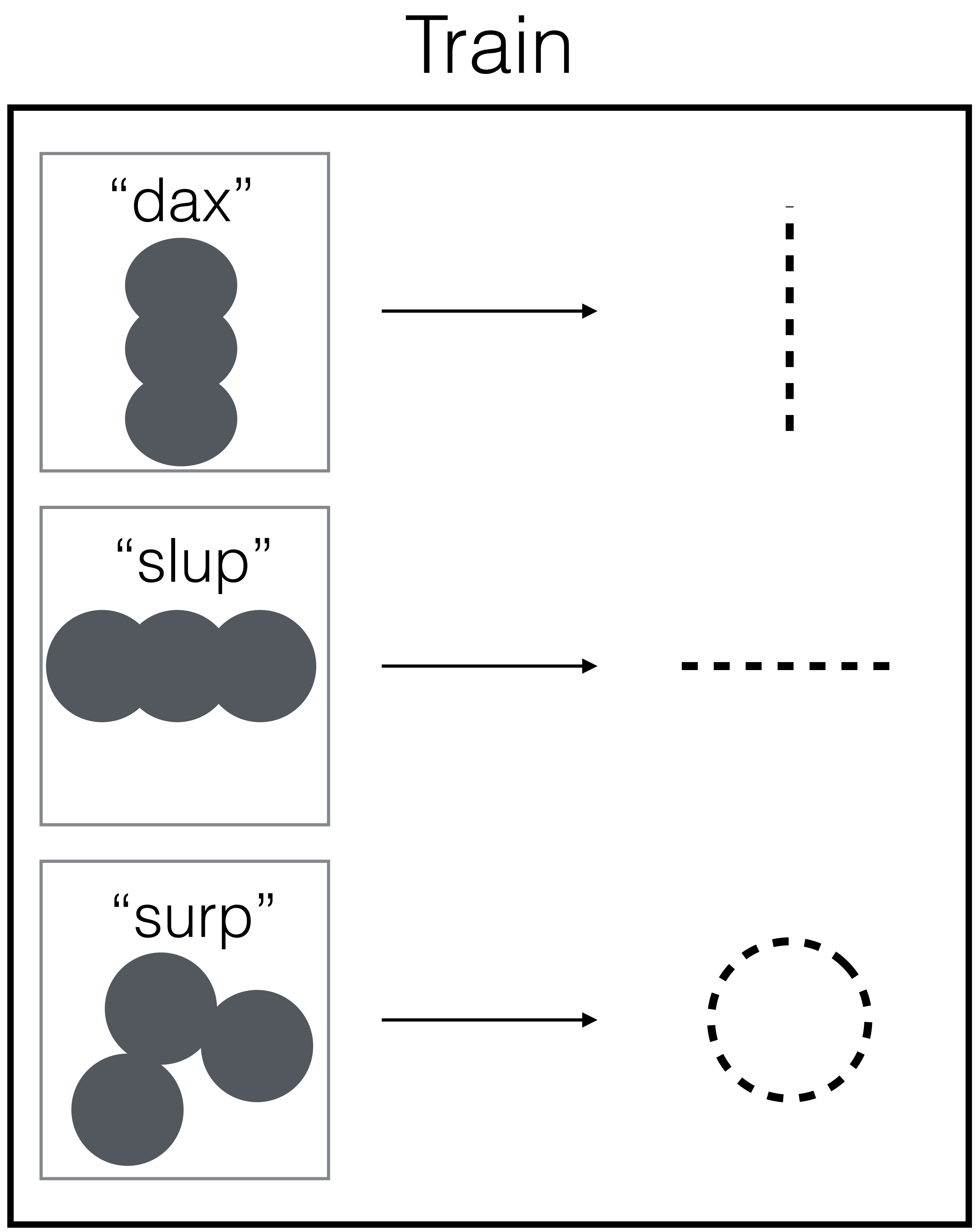
"The ability to produce/understand some sentences is *intrinsically* connected to the ability to produce/understand certain others…[they] *must be made of the same parts*."
(Fodor&Pylyshyn, 1988)

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

Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

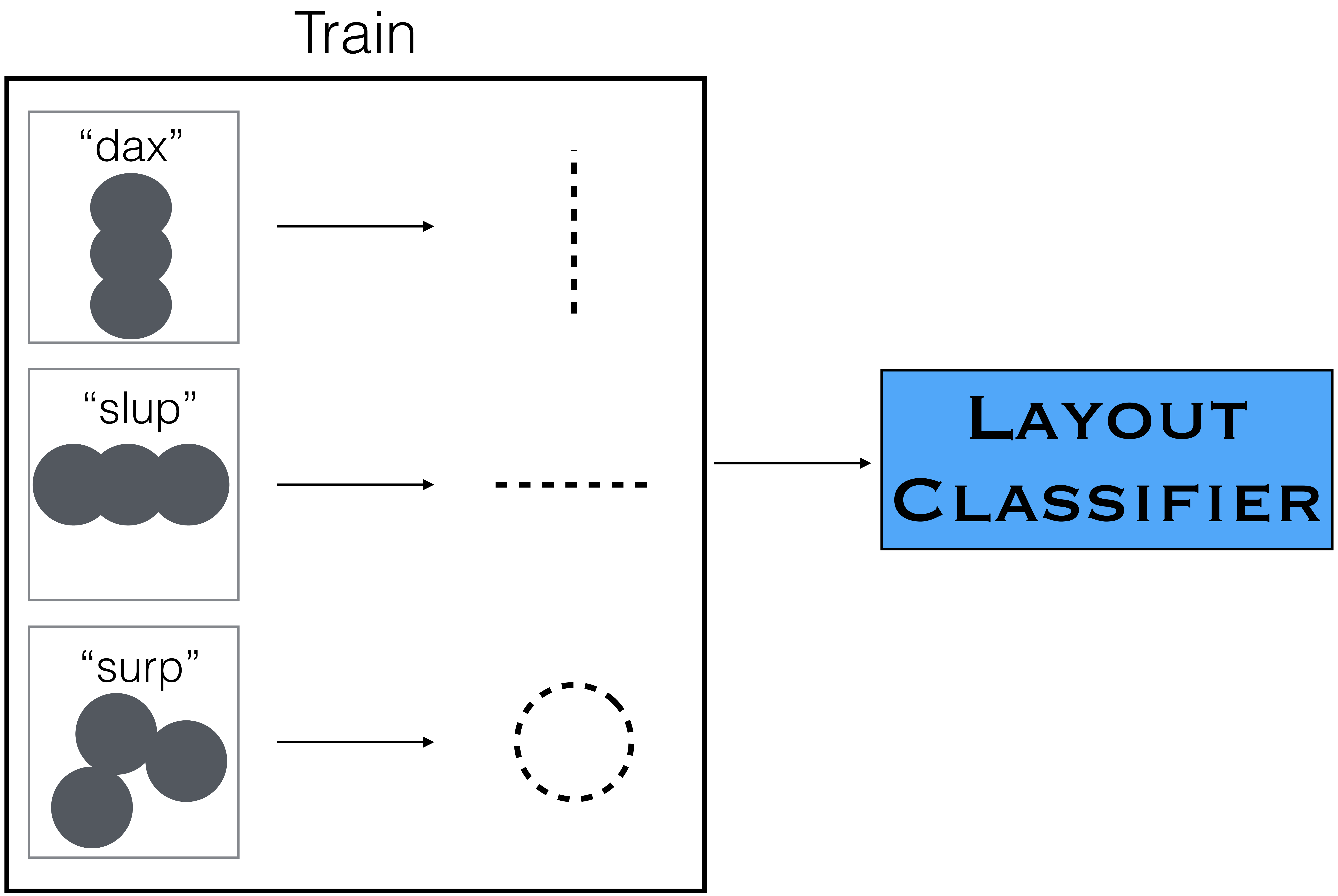
Reusability of Constituents



Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

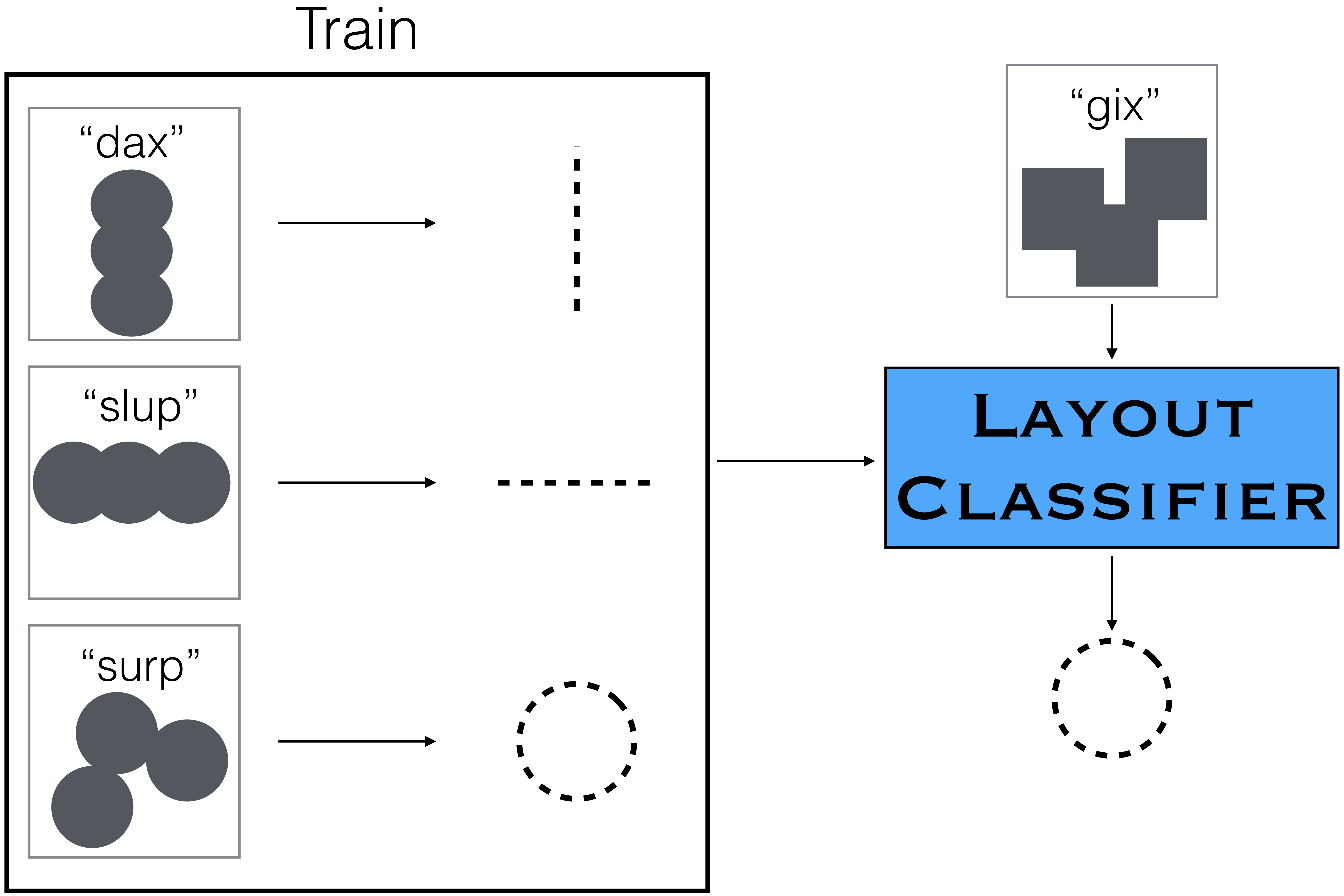
Reusability of Constituents



Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

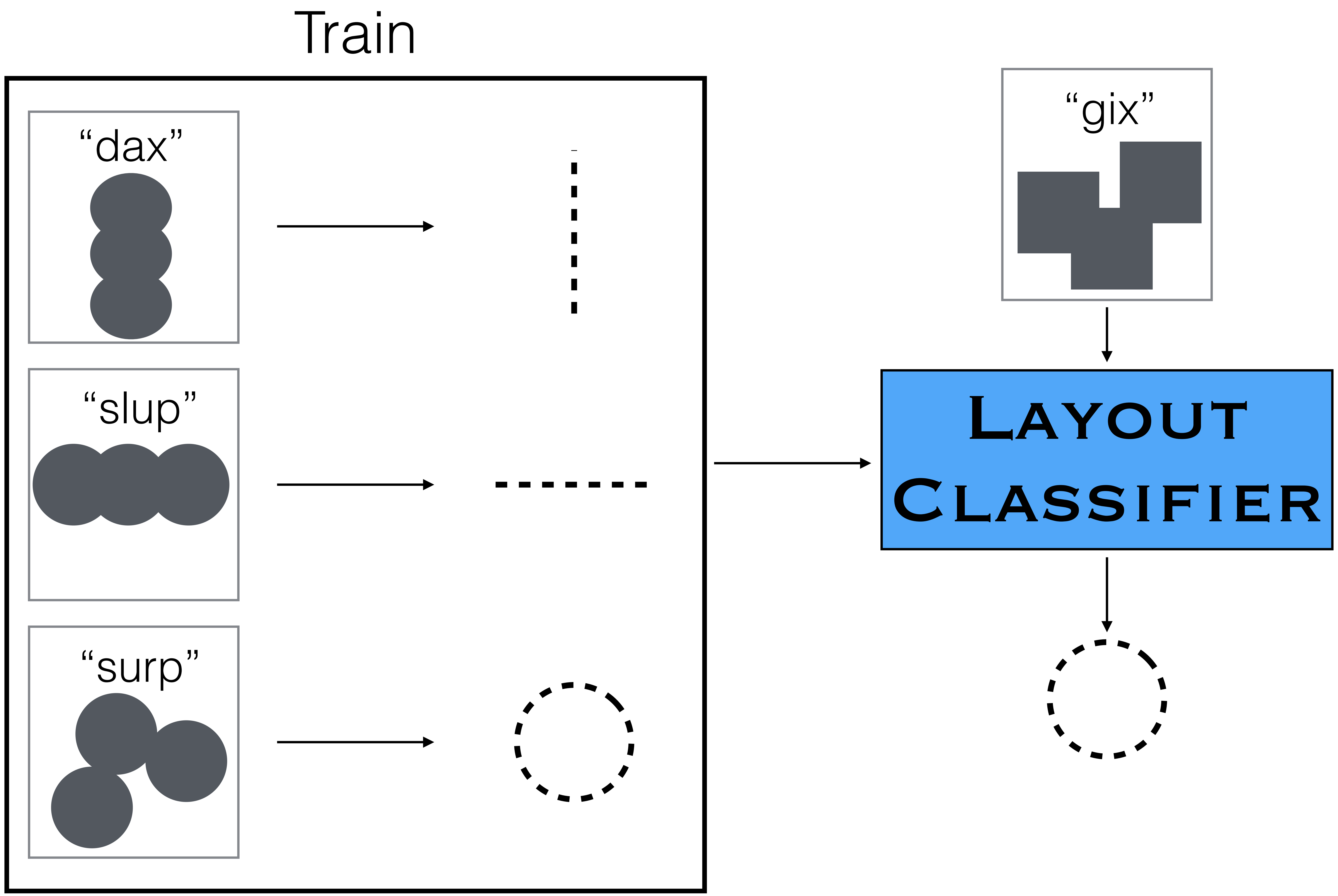
Reusability of Constituents



Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

Reusability of Constituents

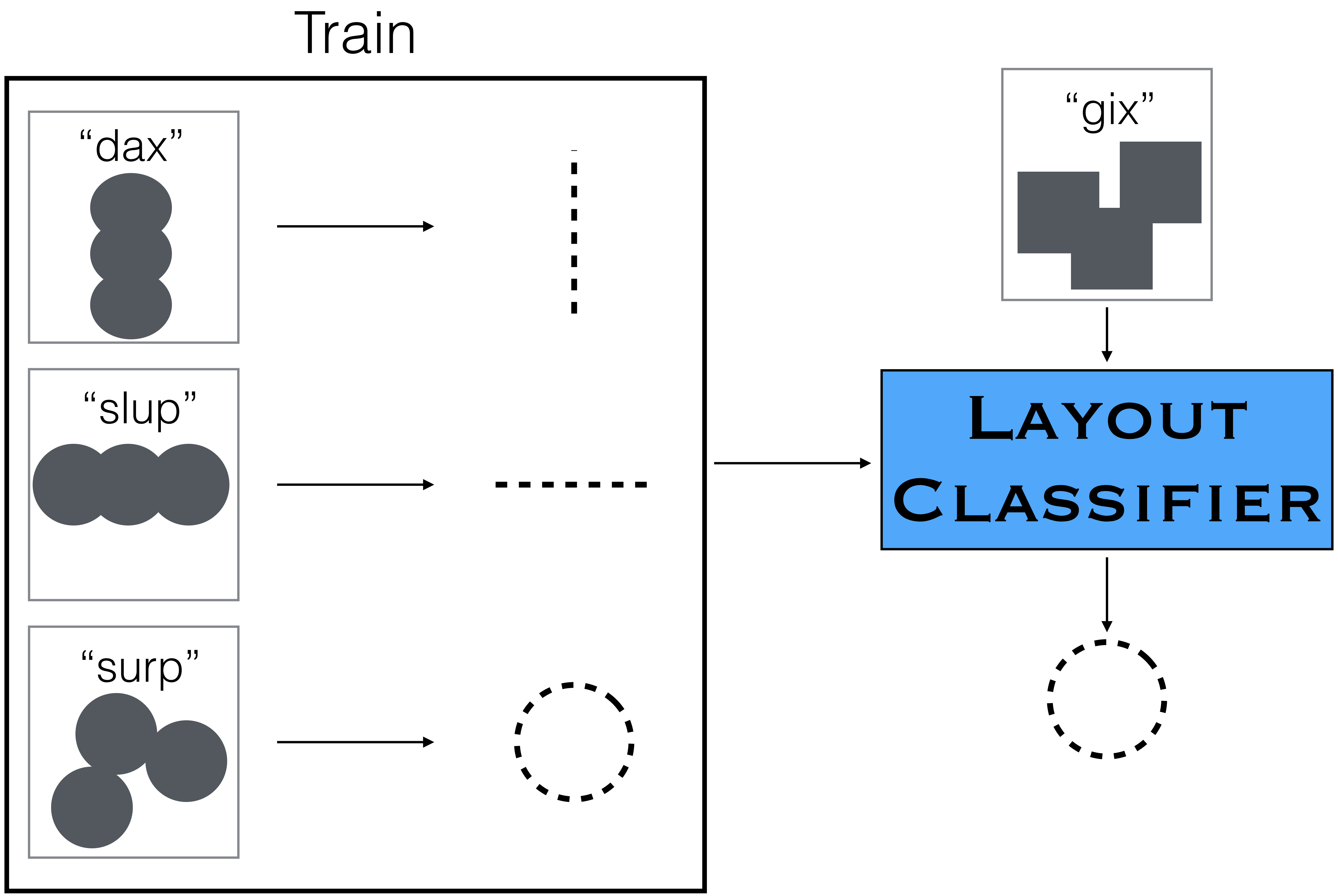


Doesn't require that the feature is discrete.

Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

Reusability of Constituents



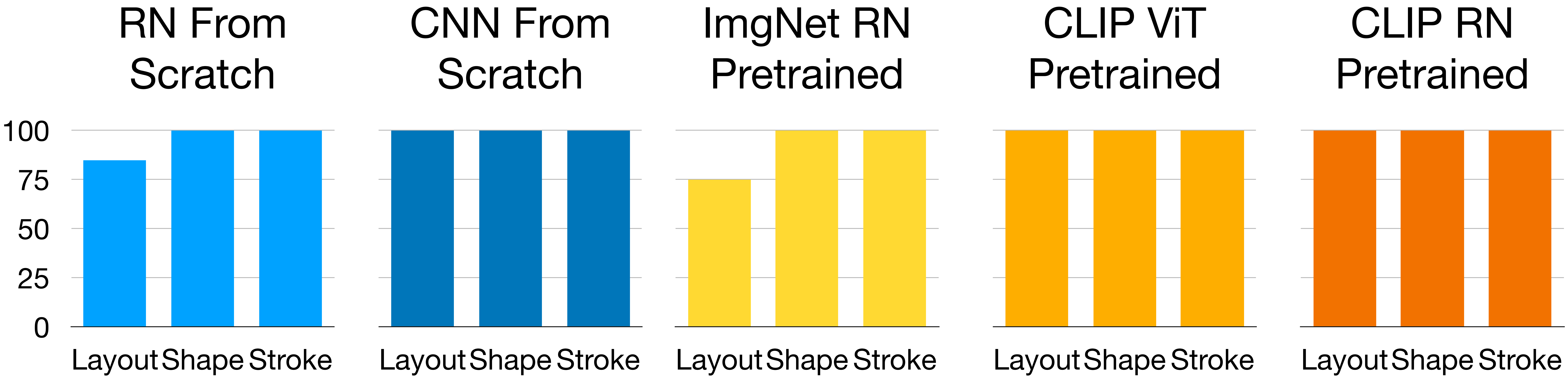
Doesn't require that the feature is discrete.

Rather, that the feature is systematically discretizable if needed.

Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

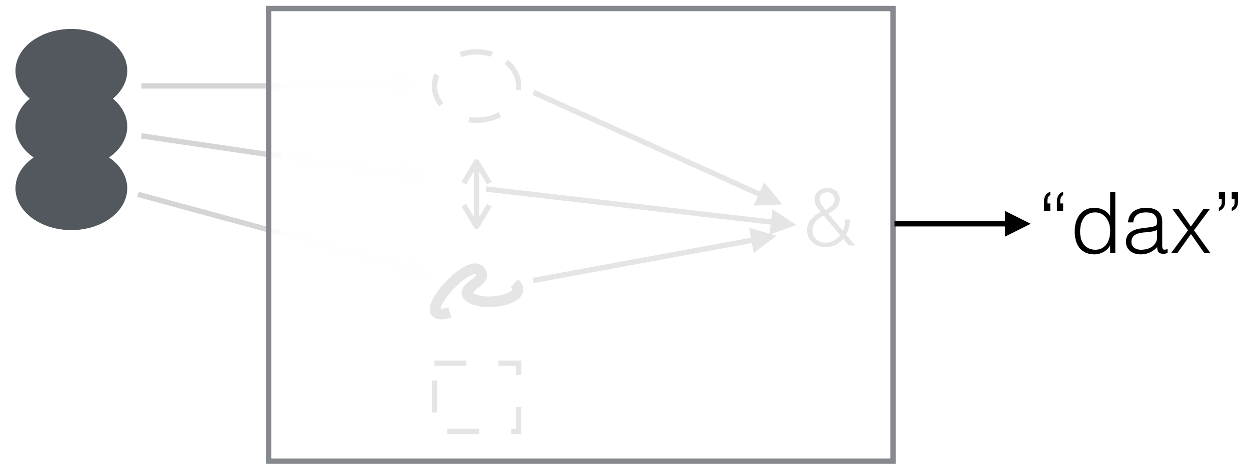
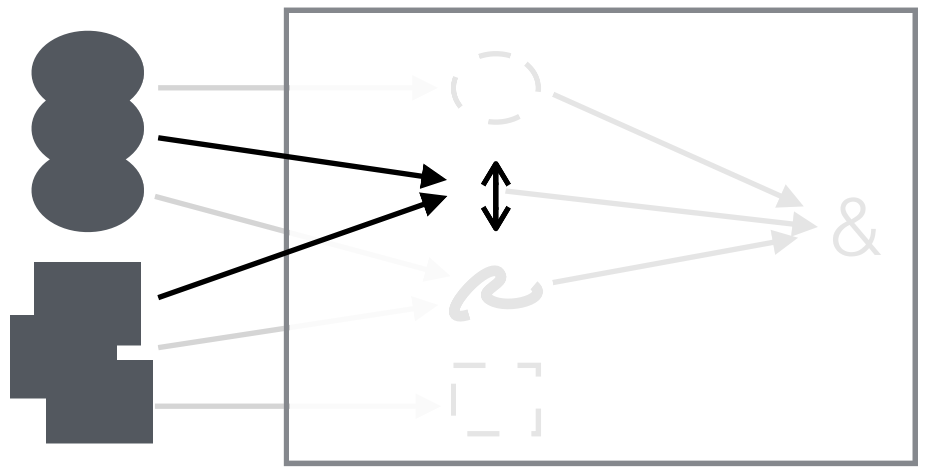
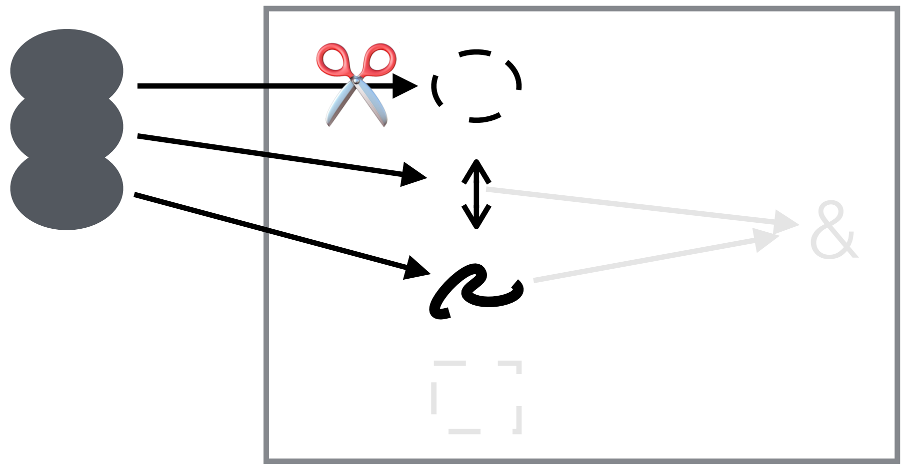
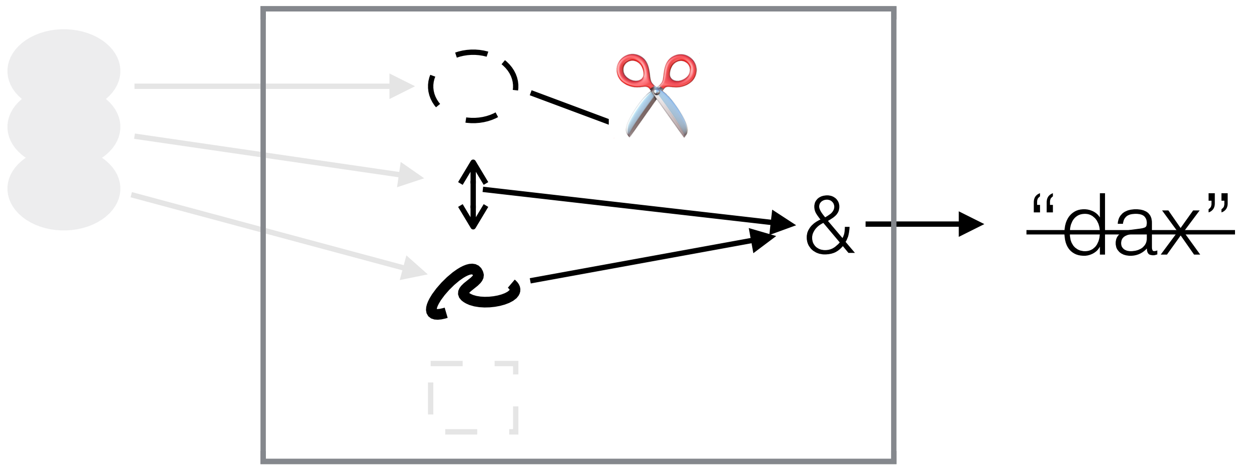
Reusability of Constituents



Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

Unit Tests

✅	Groundedness	
✅	Reusability of Constituents	
	Modularity of Constituents	
	Causality of Constituents	

Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

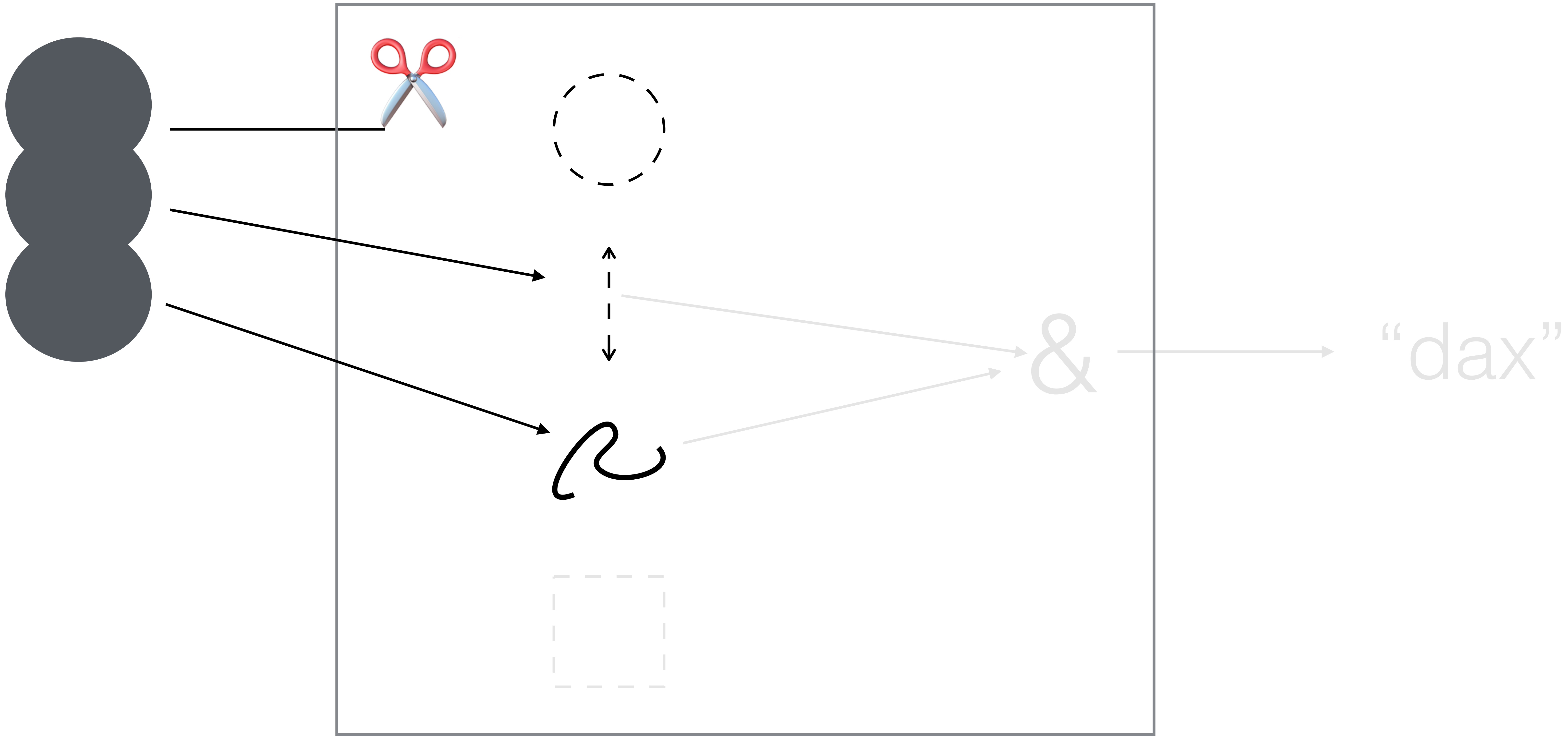
Case Study

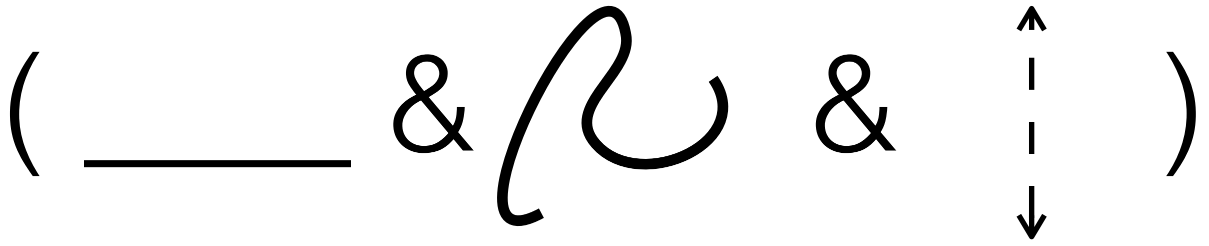
Modularity of Constituents

Case Study

Modularity of Constituents

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

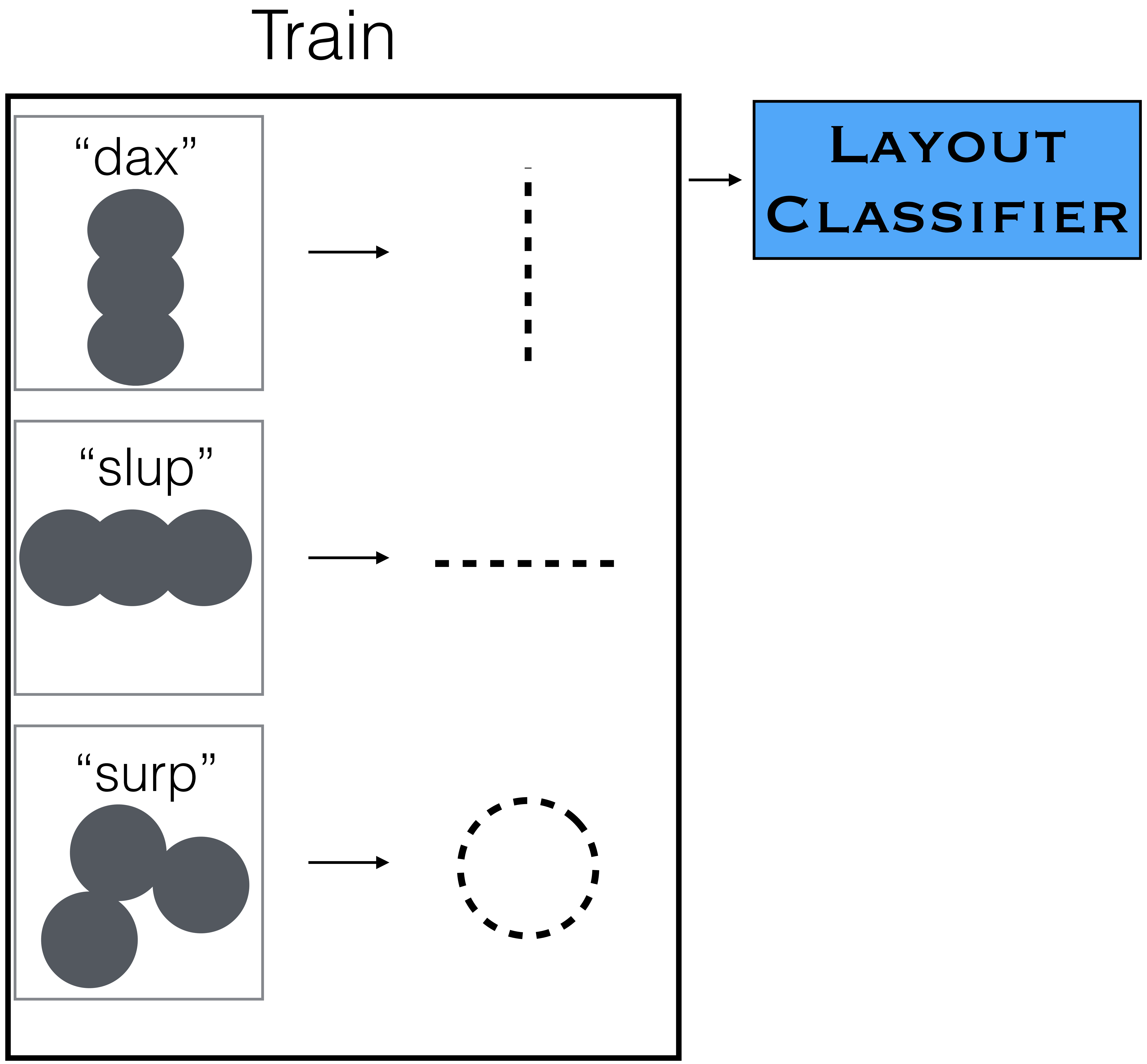




Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

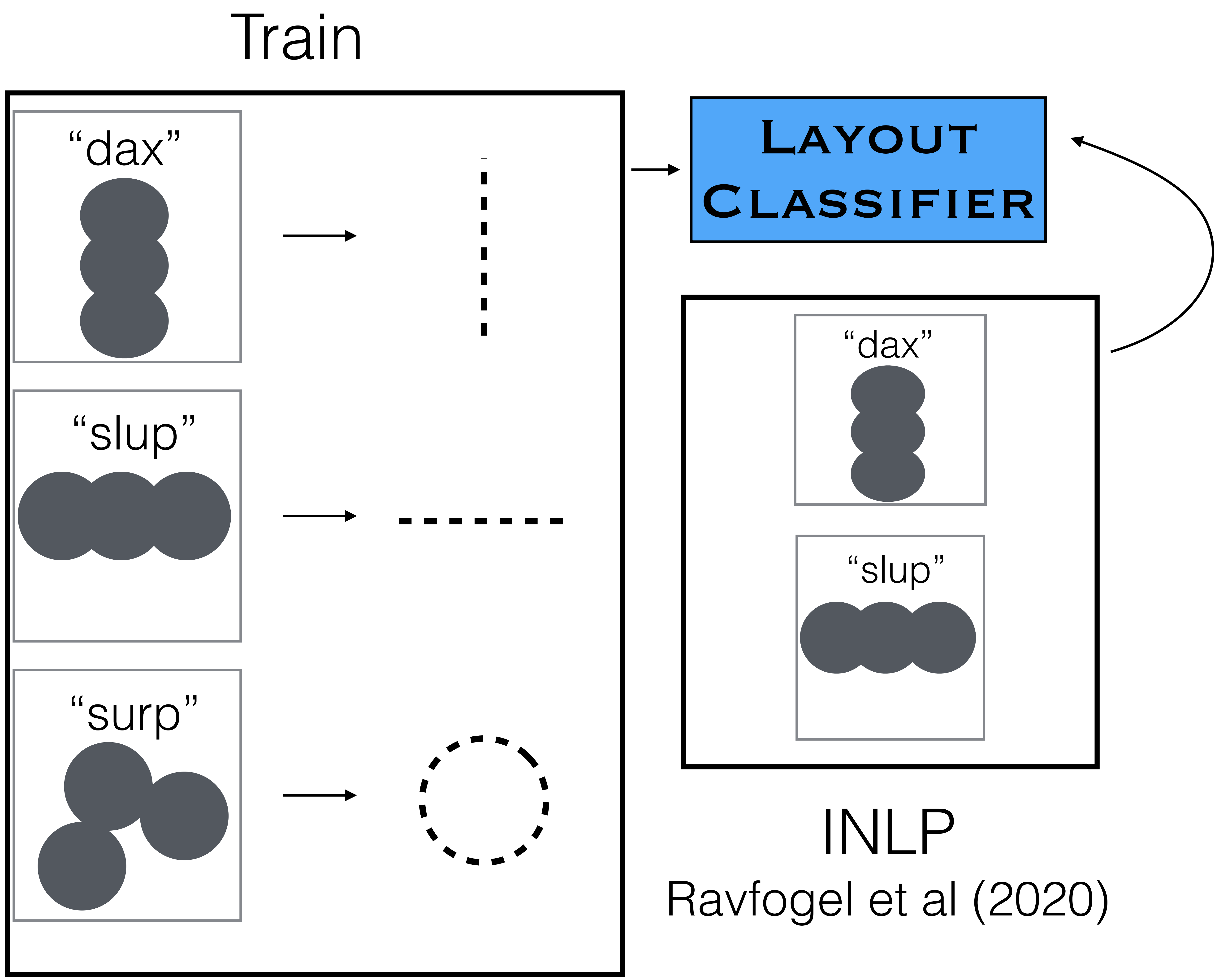
Modularity of Constituents



Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

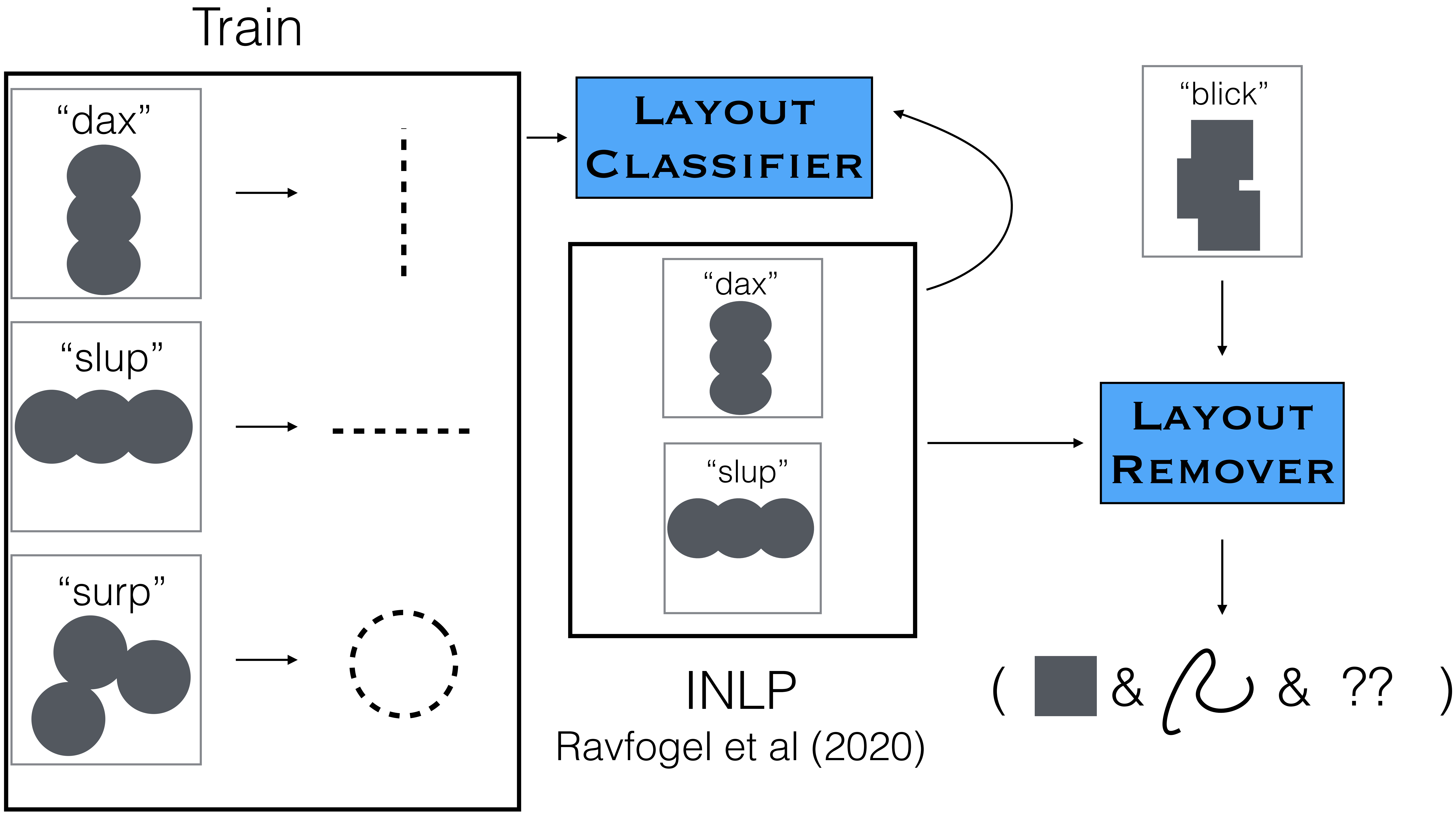
Modularity of Constituents



Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

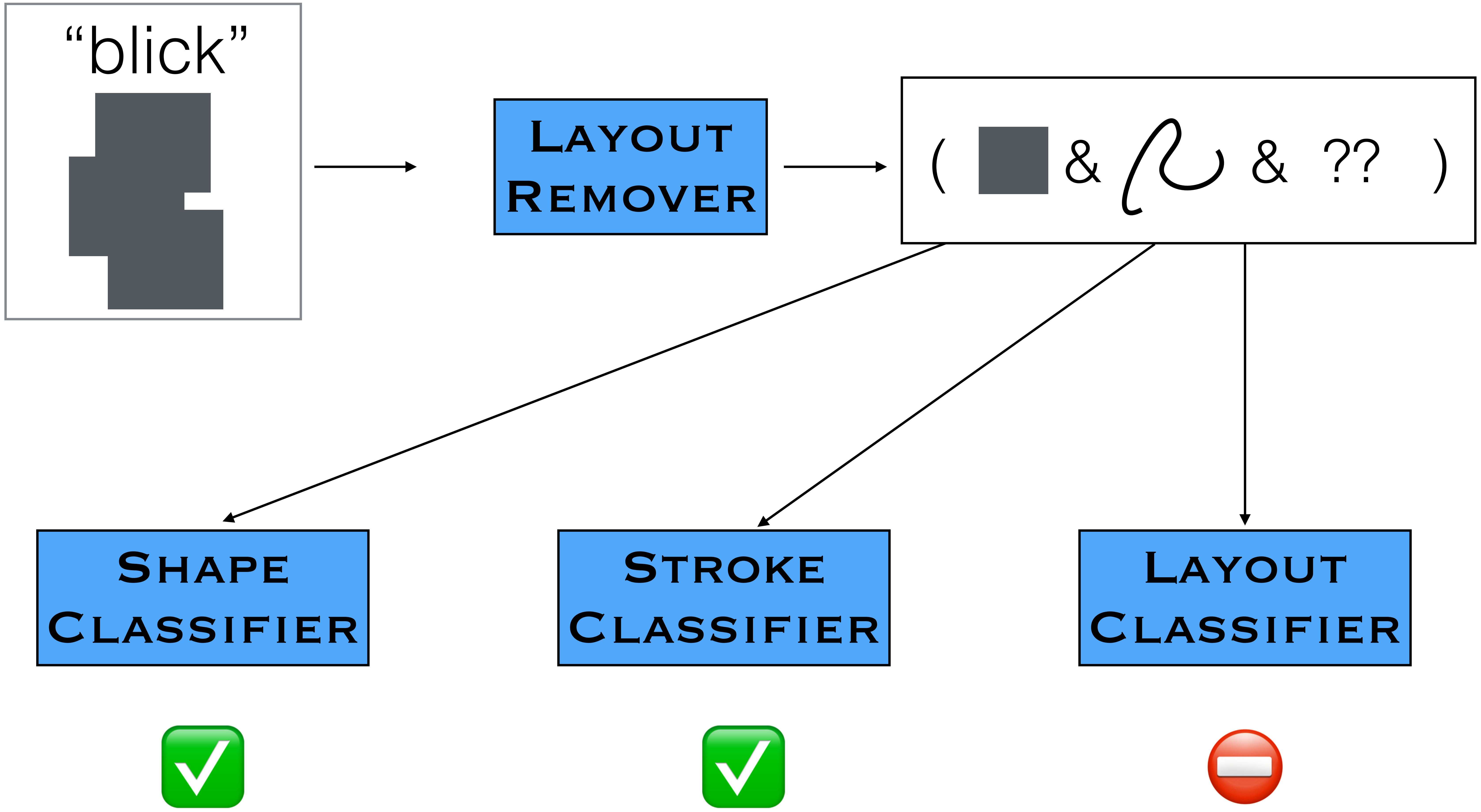
Case Study

Modularity of Constituents



Case Study

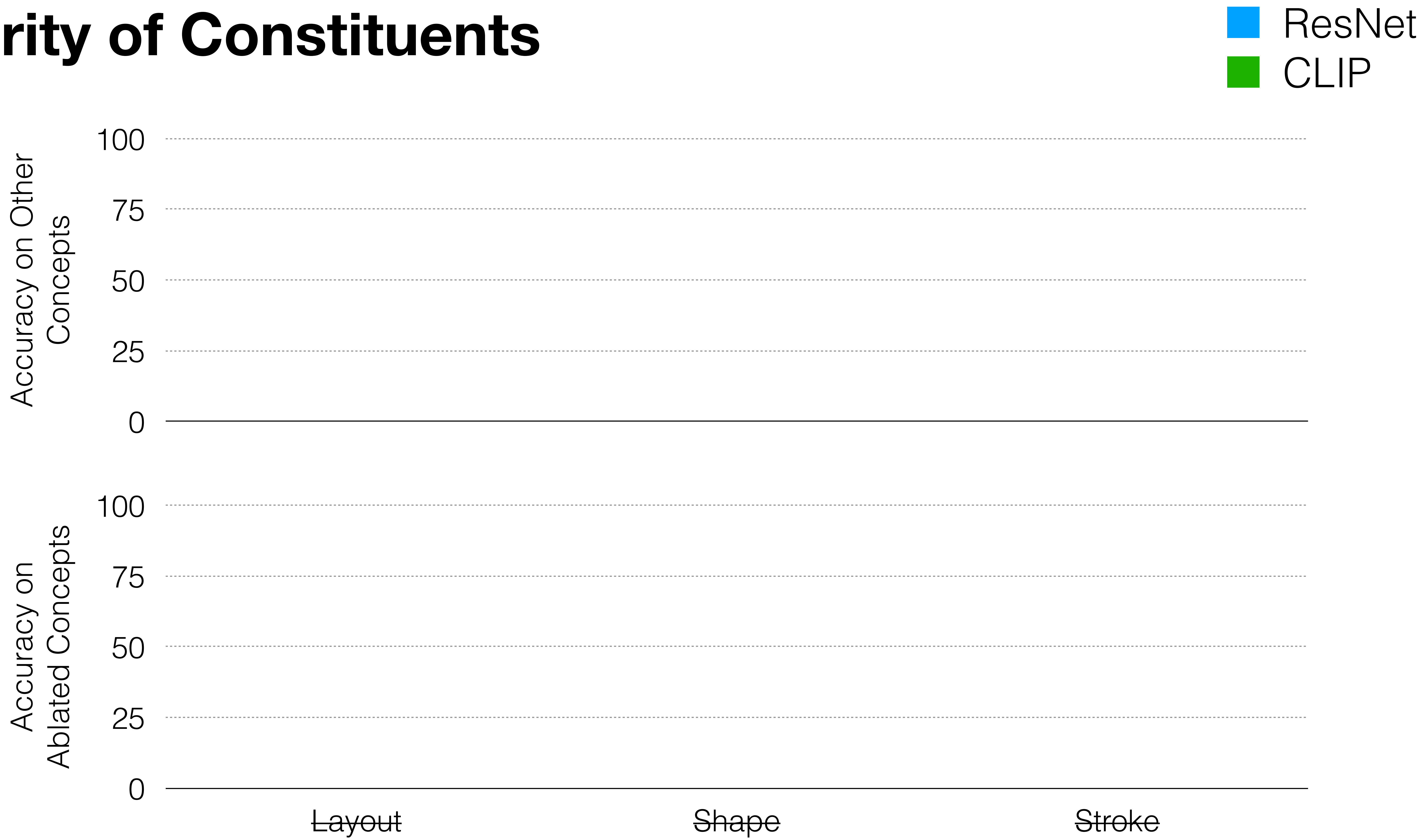
Modularity of Constituents



Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

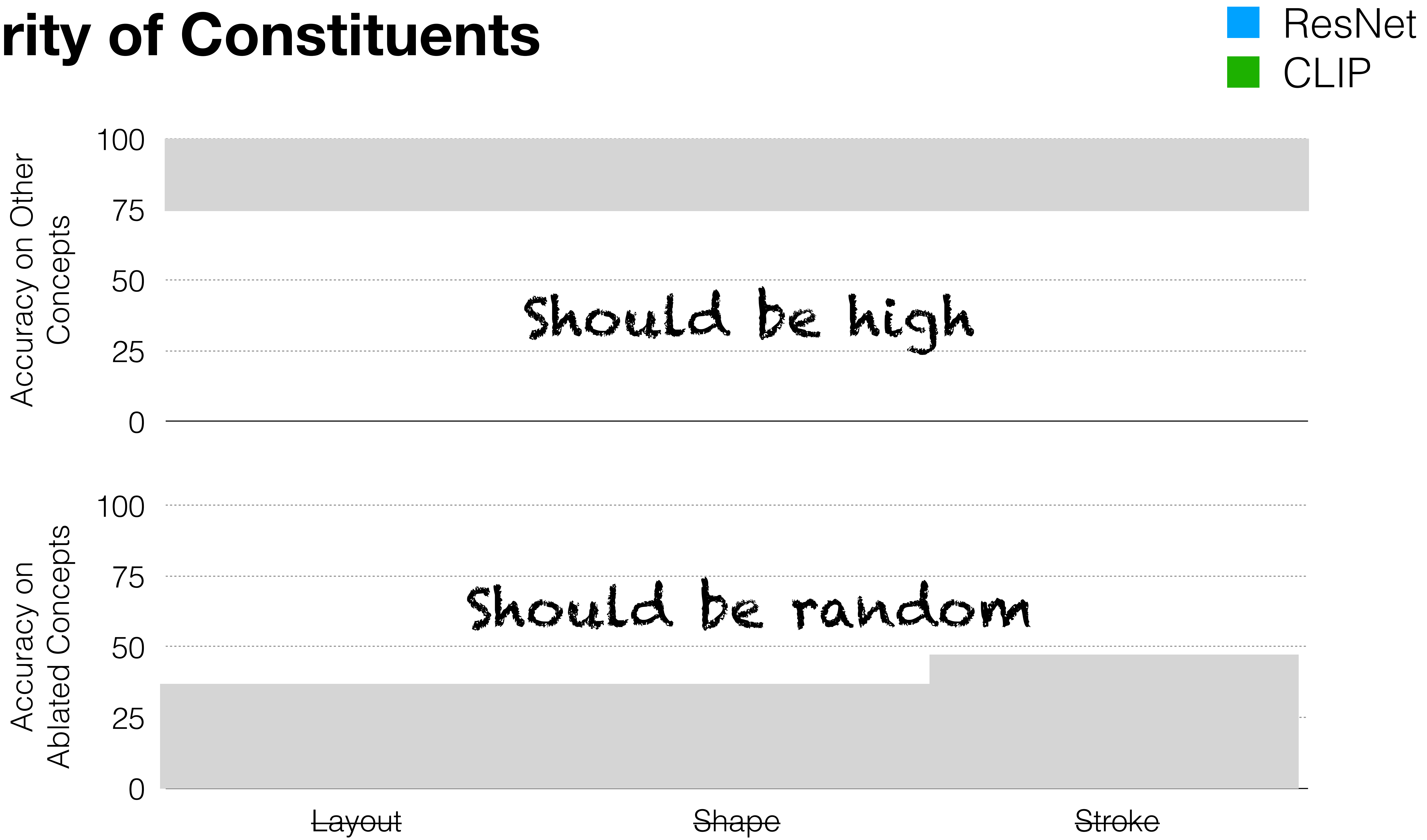
Modularity of Constituents



Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

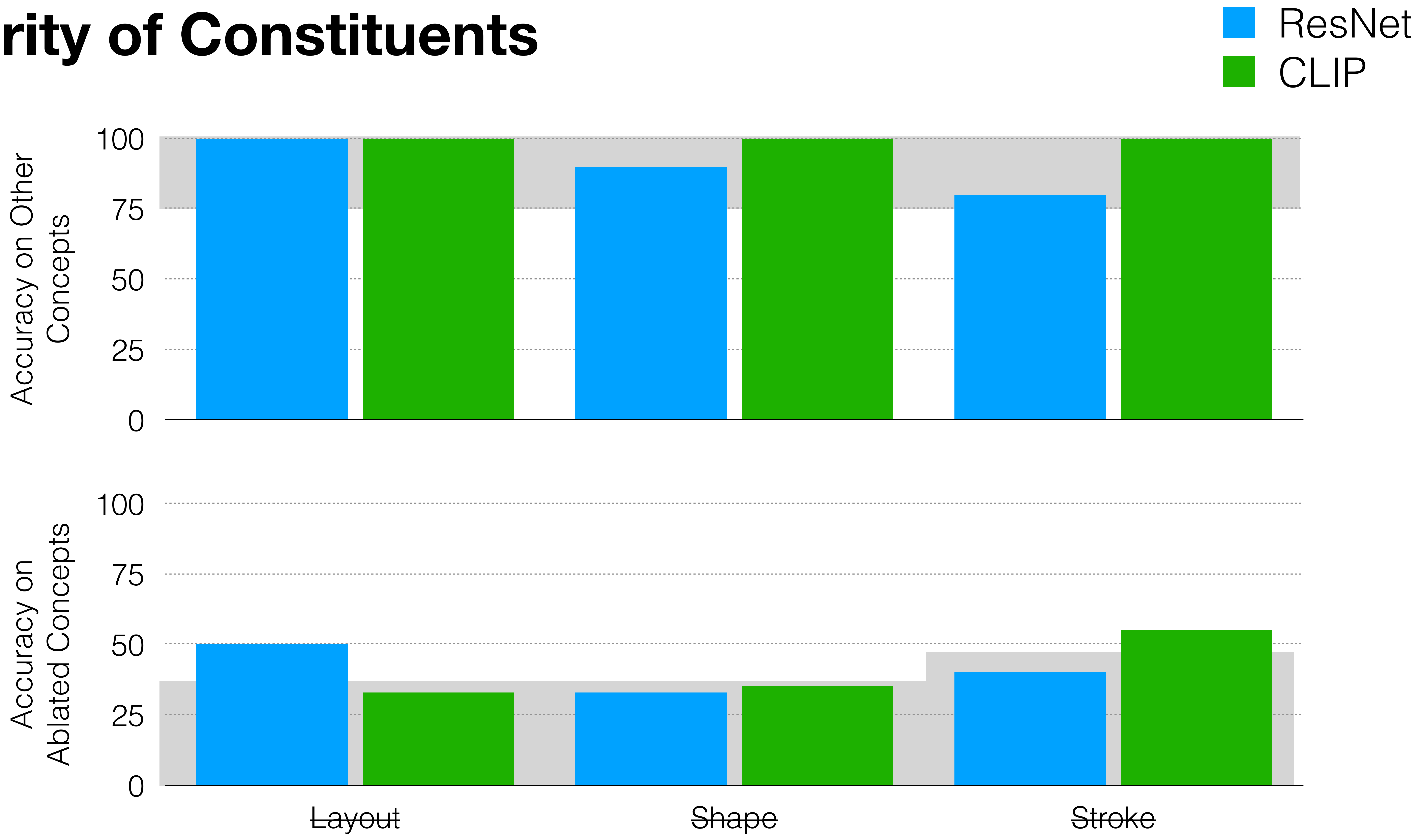
Modularity of Constituents



Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

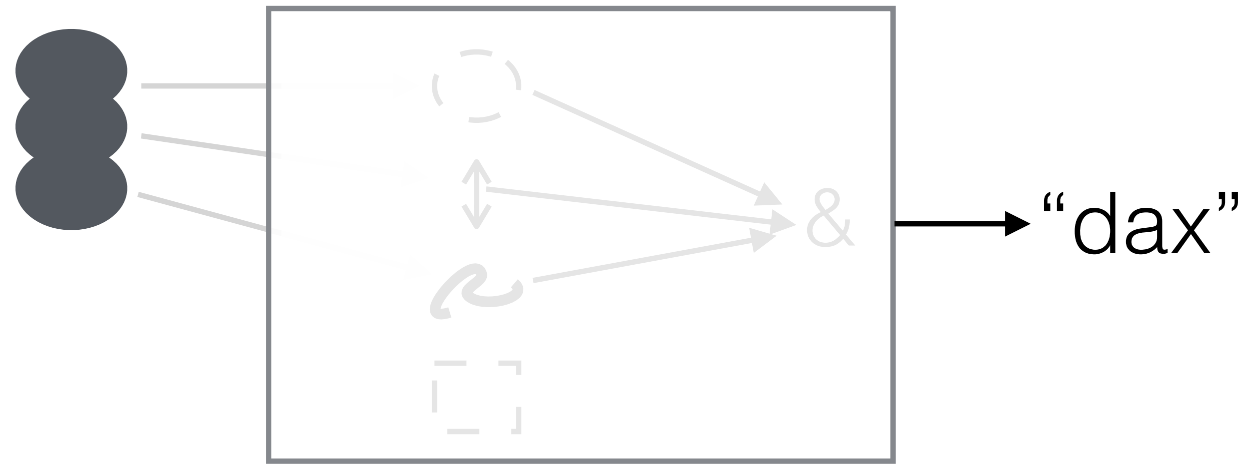
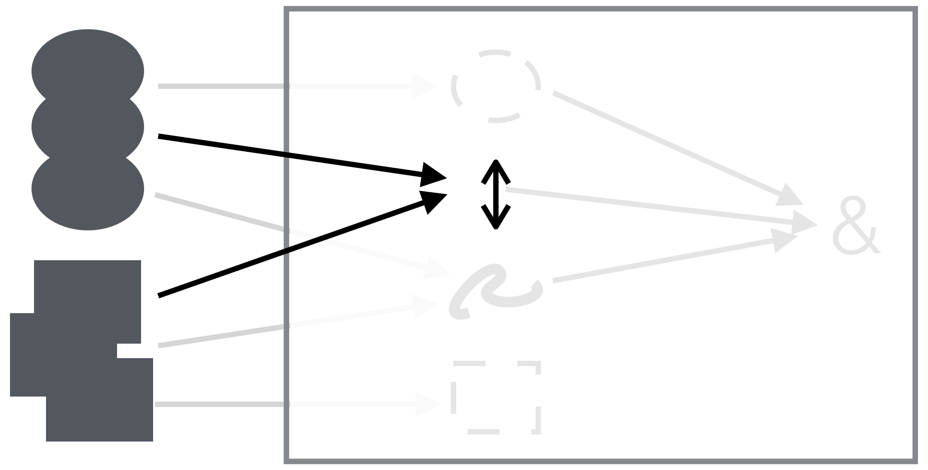
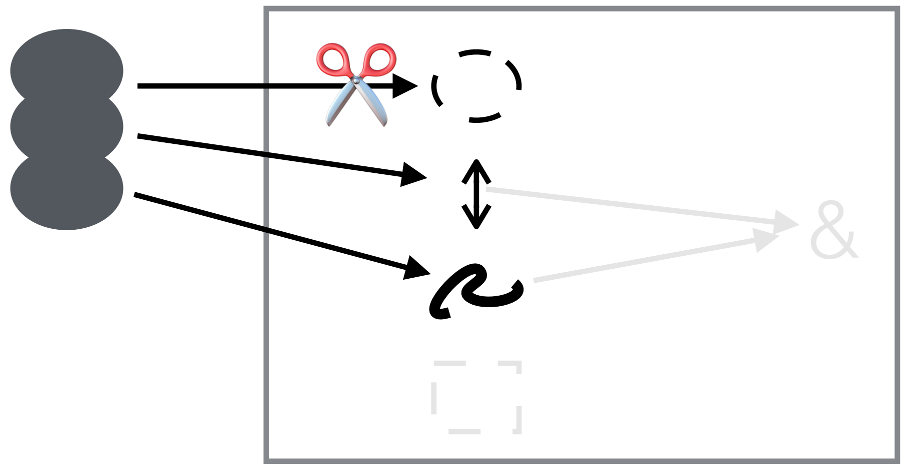
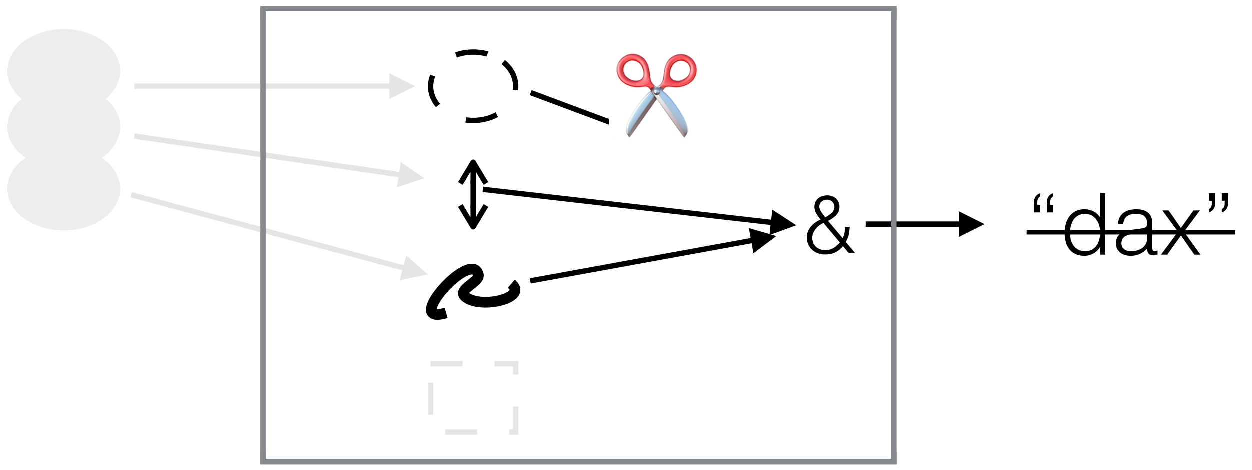
Modularity of Constituents



Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

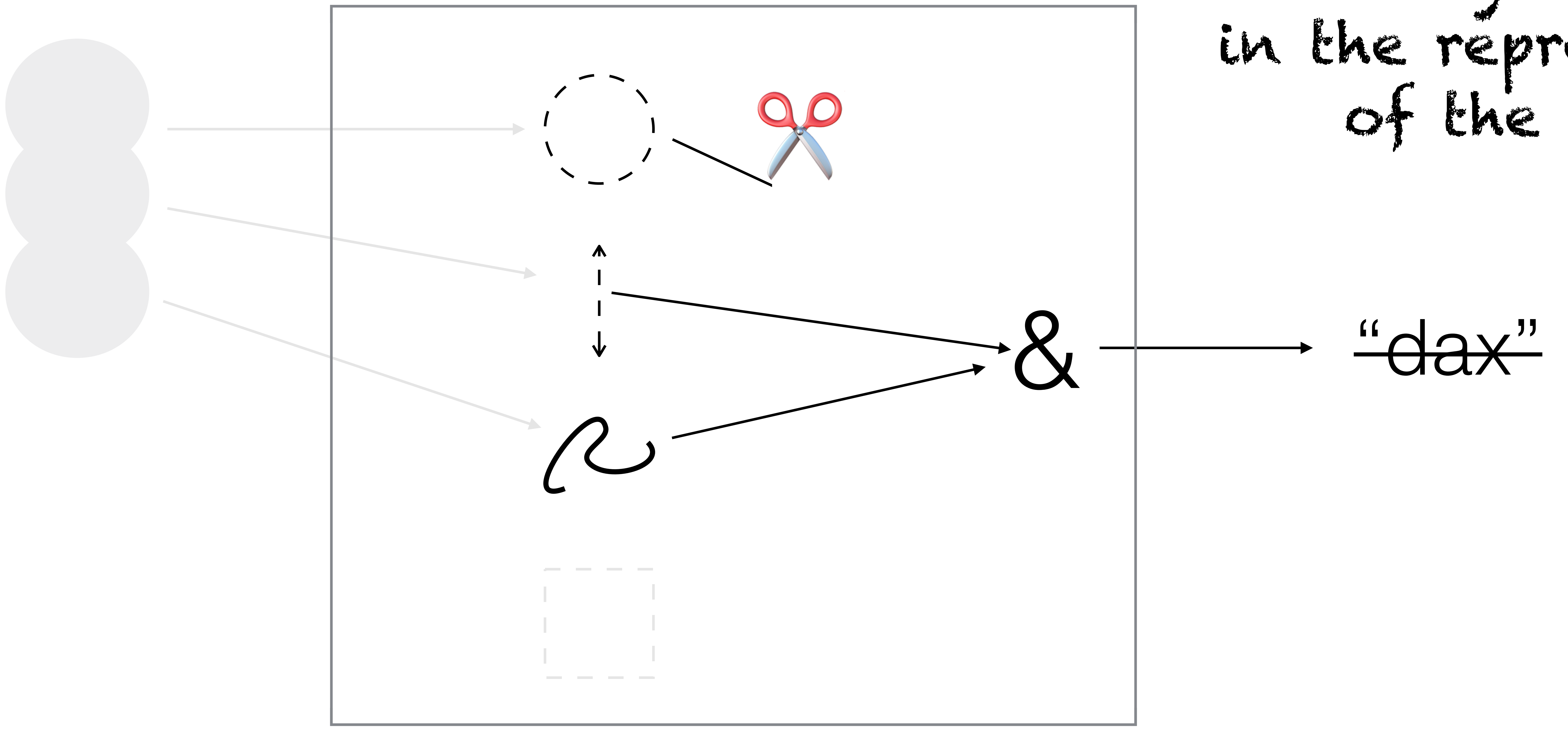
Unit Tests

✅	Groundedness	
✅	Reusability of Constituents	
✅	Modularity of Constituents	
	Causality of Constituents	

Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

Causality of Constituents



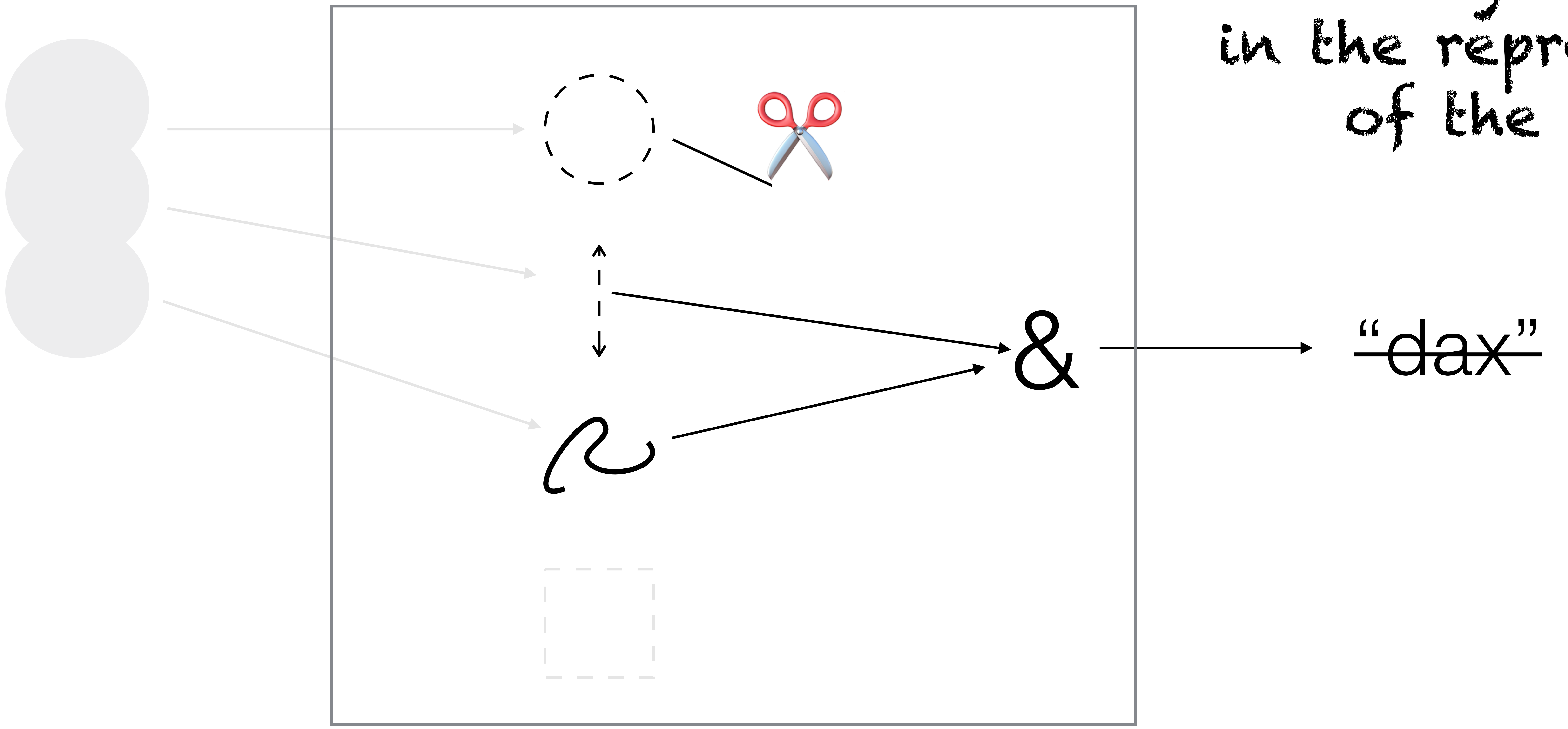
Representations of the parts are causally implicated in the representation of the whole.

Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

Causality of Constituents

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

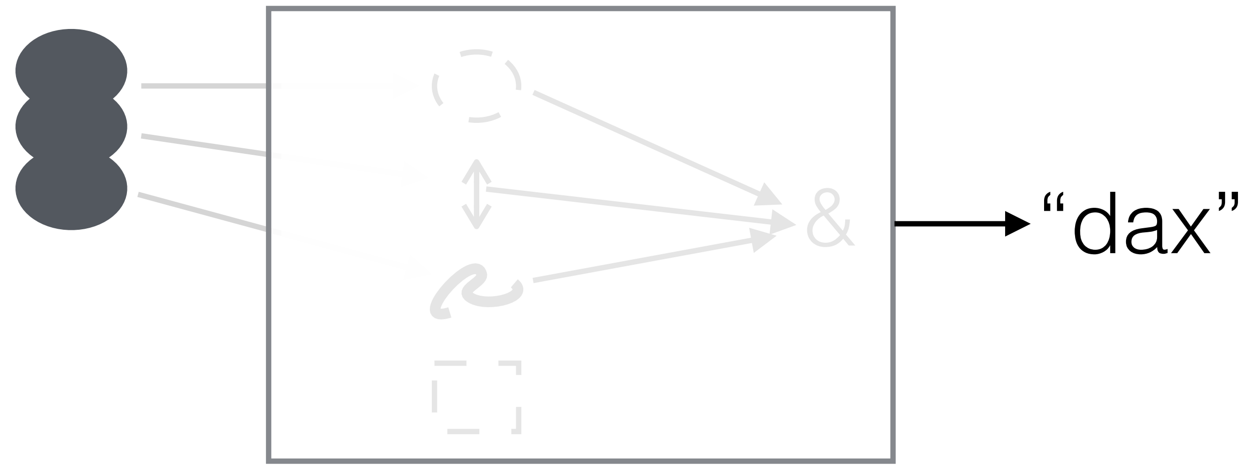
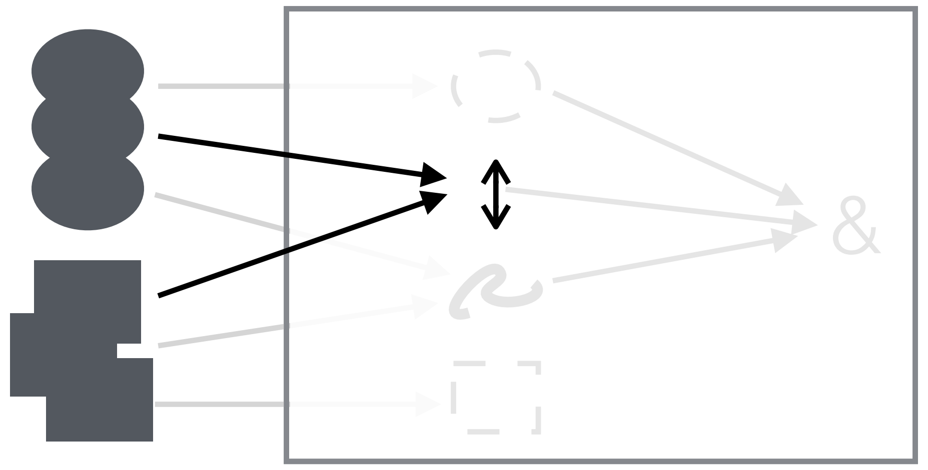
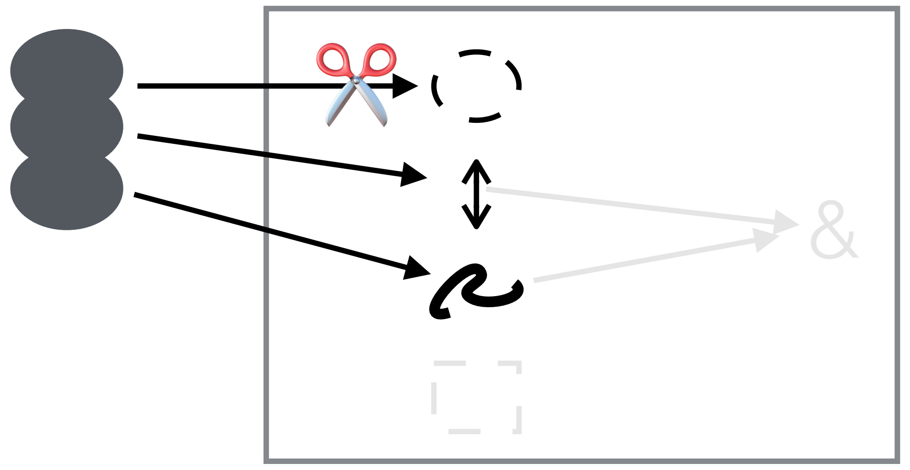
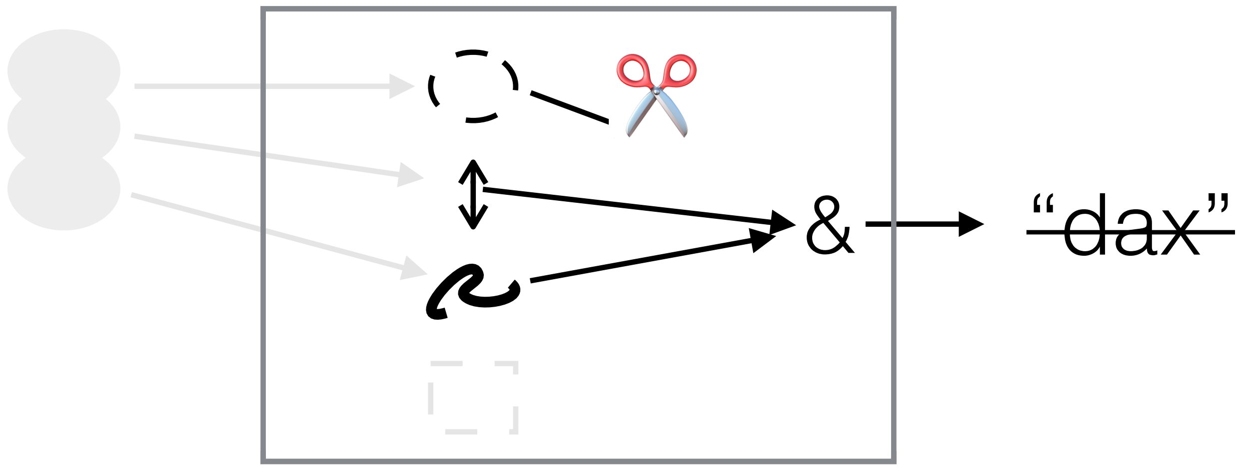


Results on this test are very much in progress, and very much in the weeds.

Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

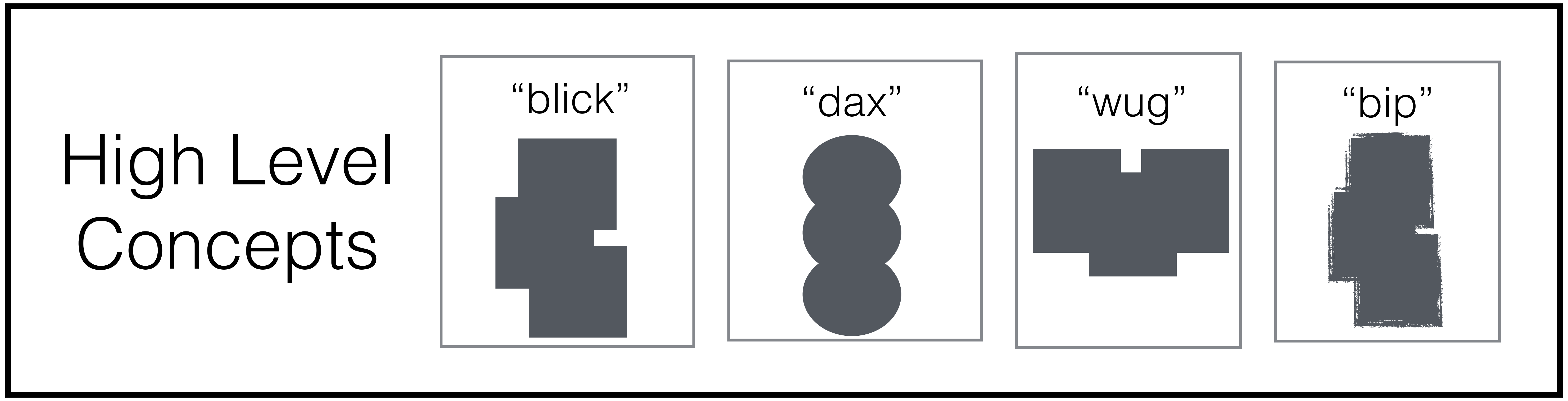
Unit Tests

✅	Groundedness	
✅	Reusability of Constituents	
✅	Modularity of Constituents	
⚠️	Causality of Constituents	

Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

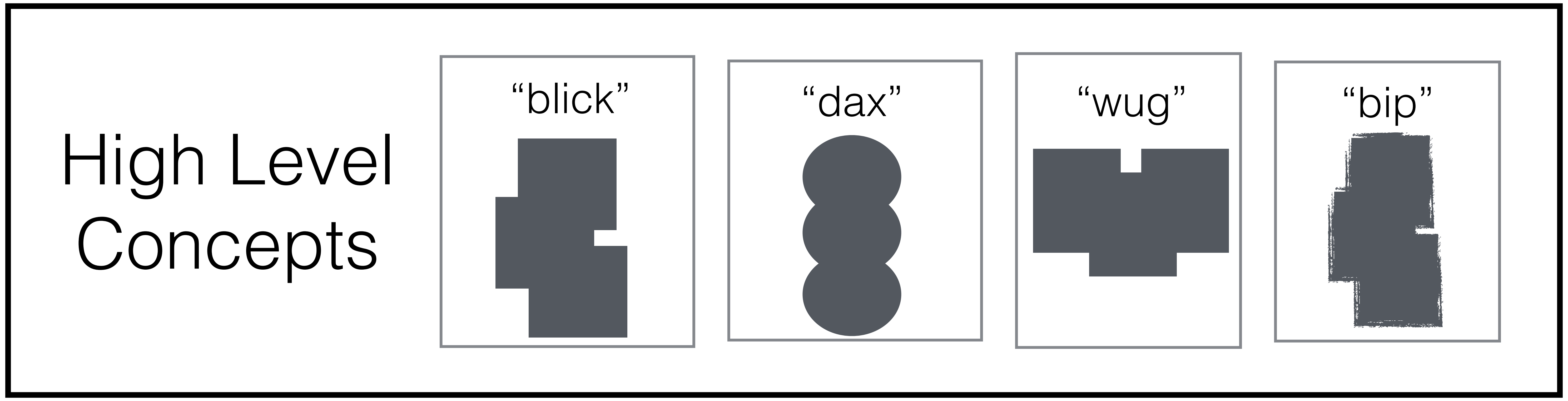
Causality of Constituents

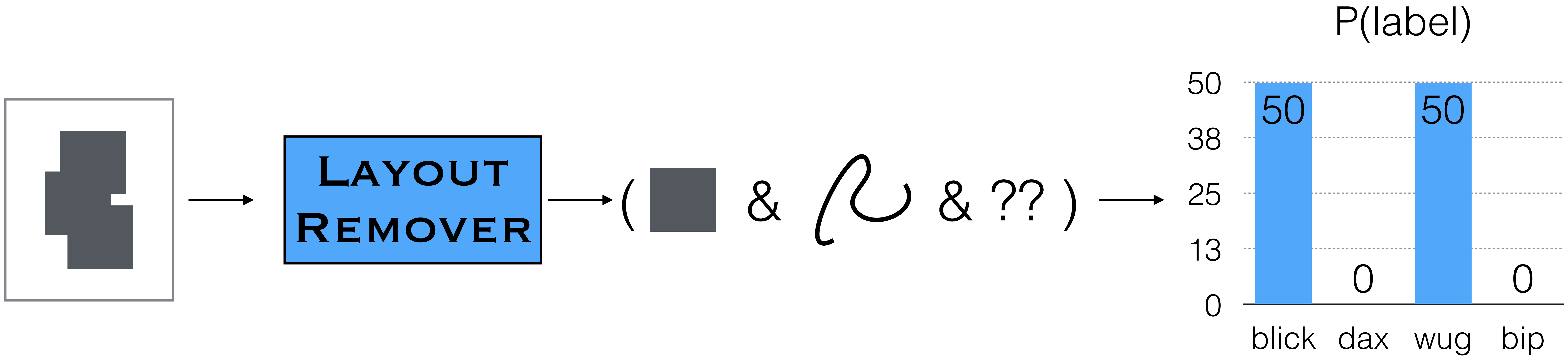


Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

Causality of Constituents

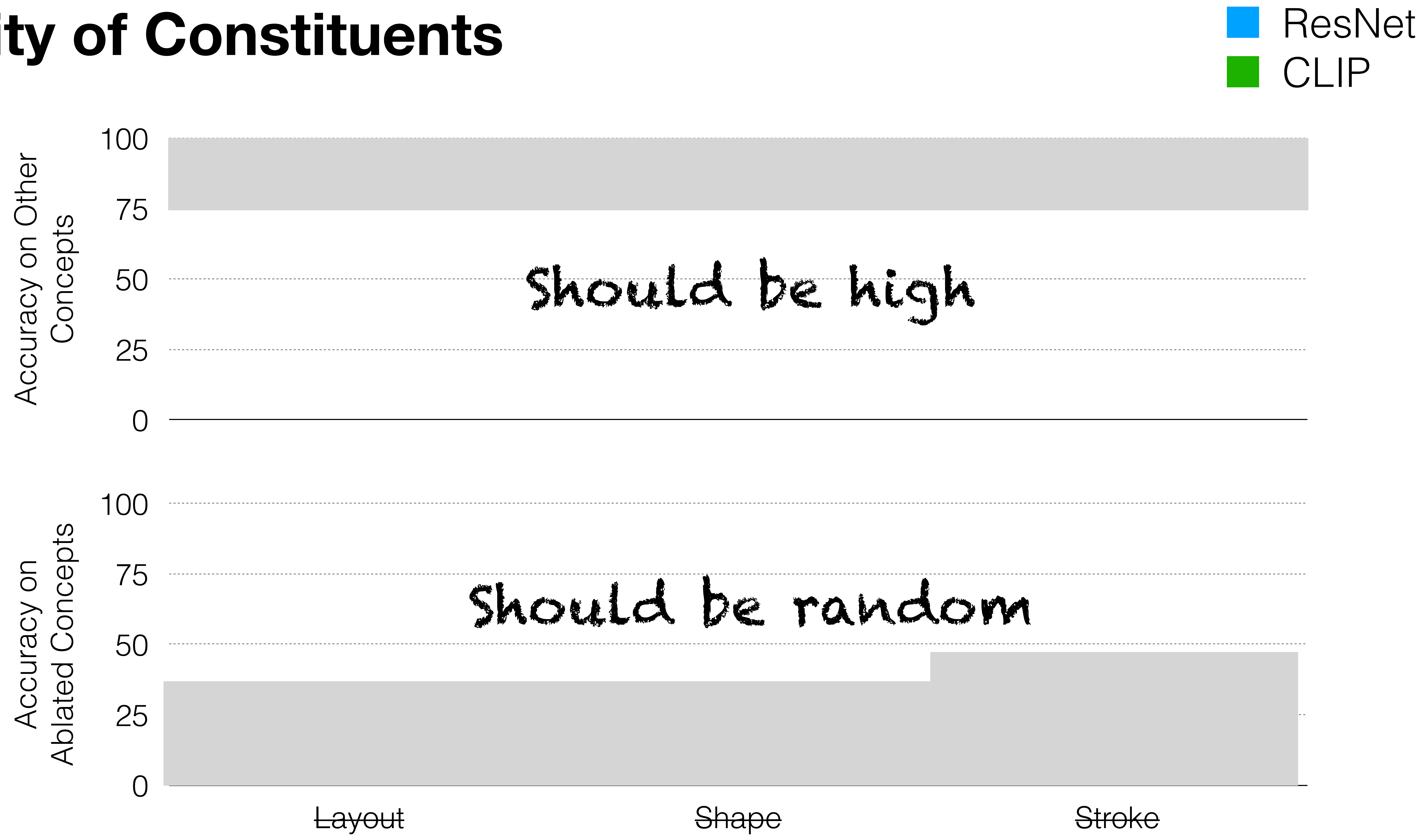




Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

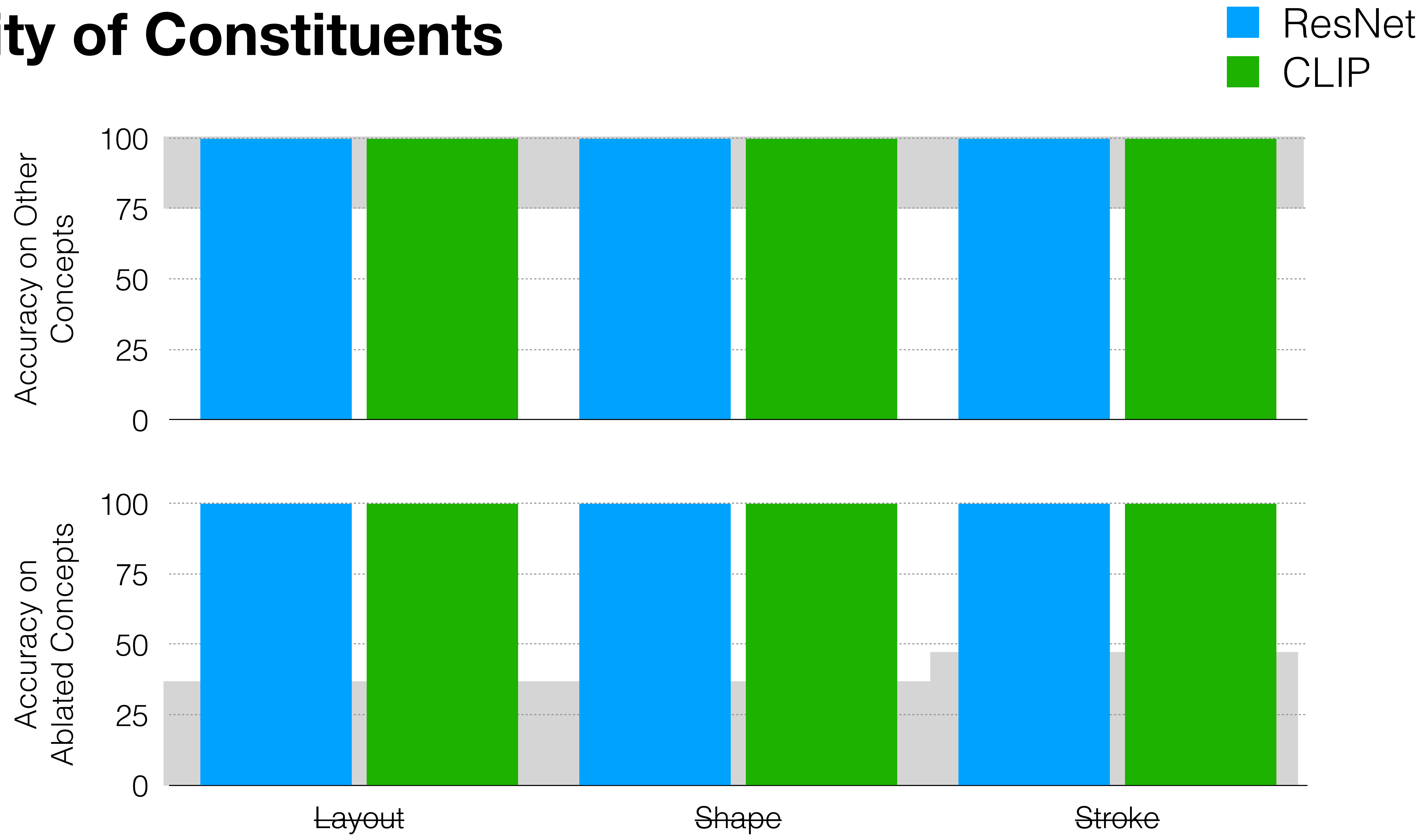
Causality of Constituents



Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

Causality of Constituents



Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

Causality of Constituents

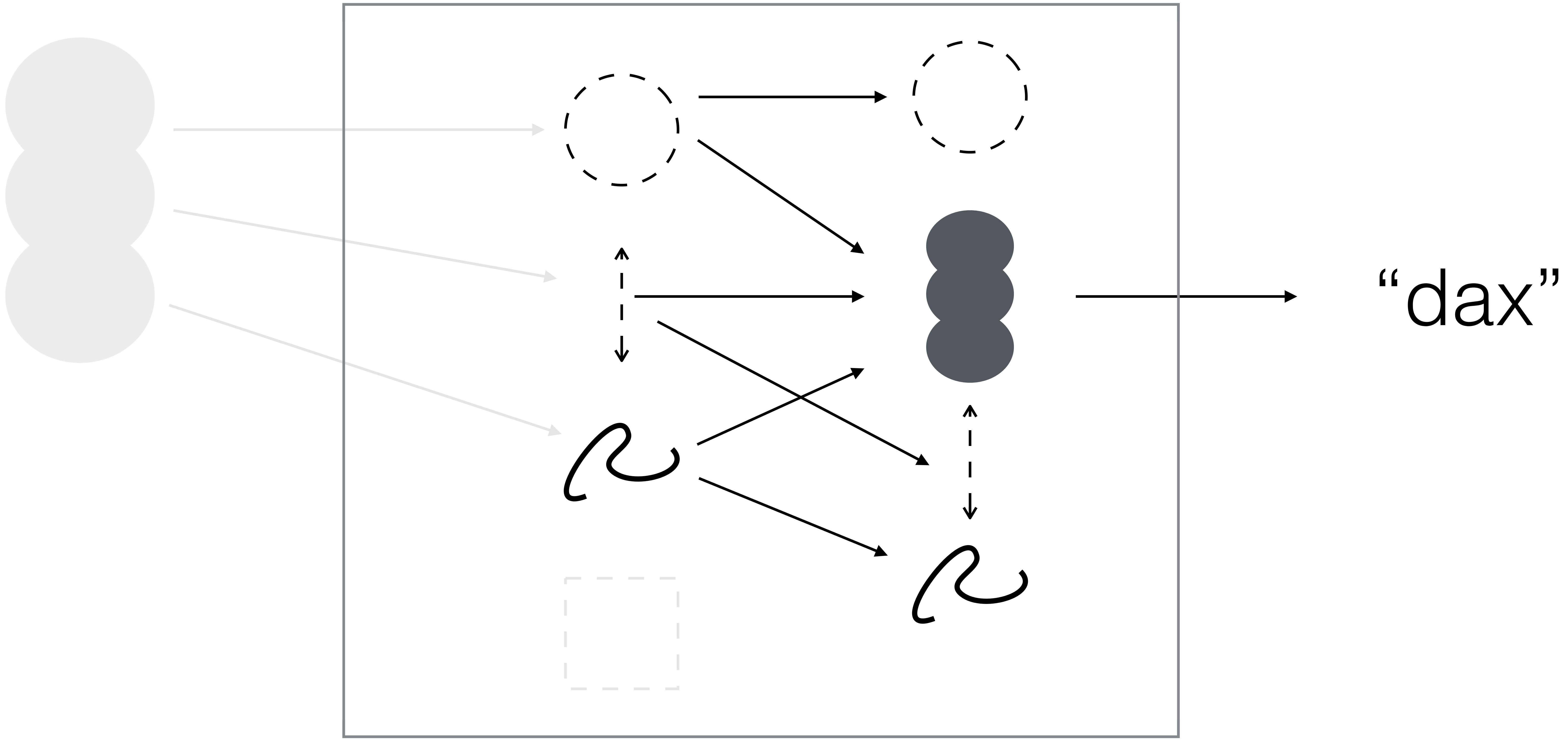
Composition across layers?

Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

Causality of Constituents

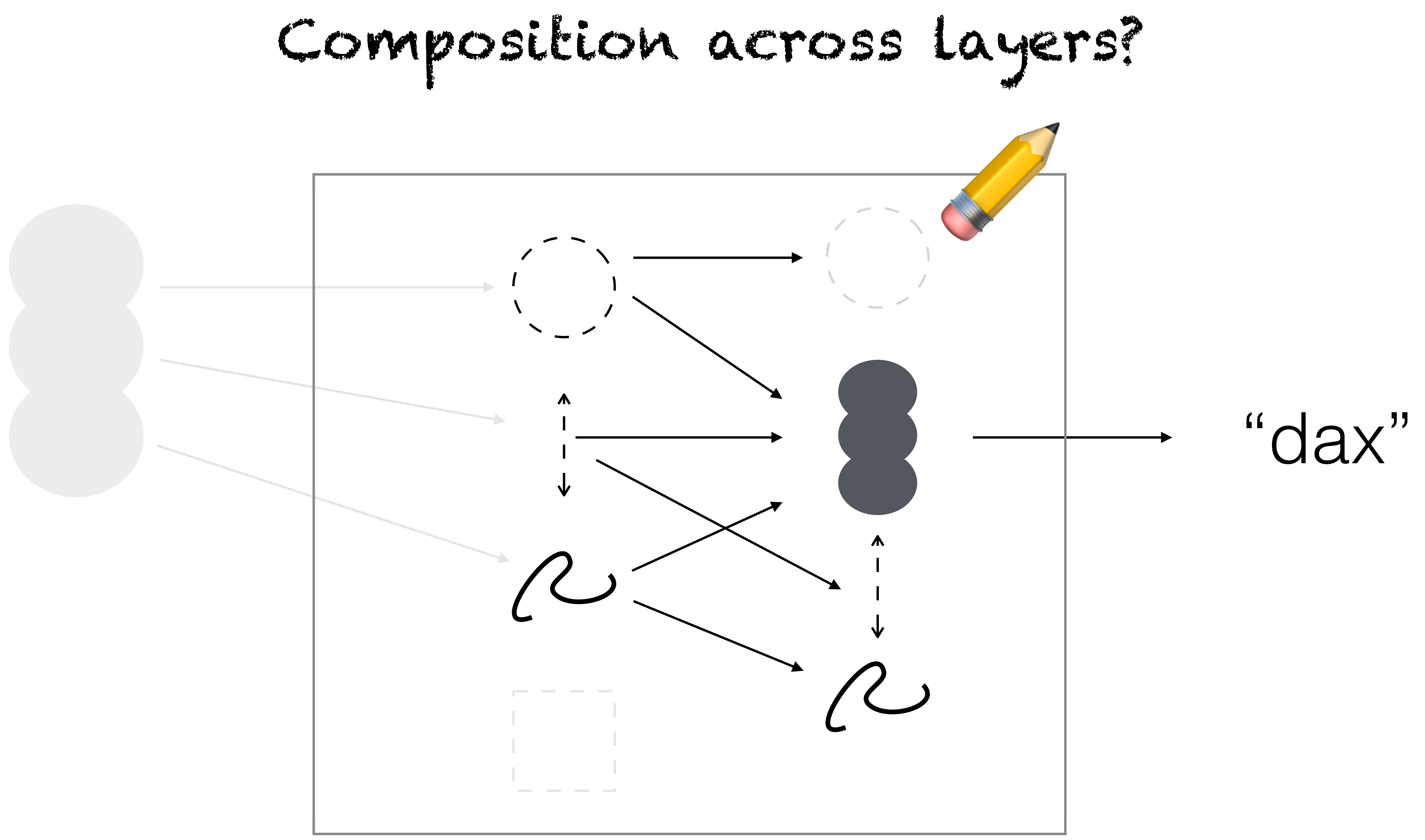
Composition across layers?



Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

Causality of Constituents

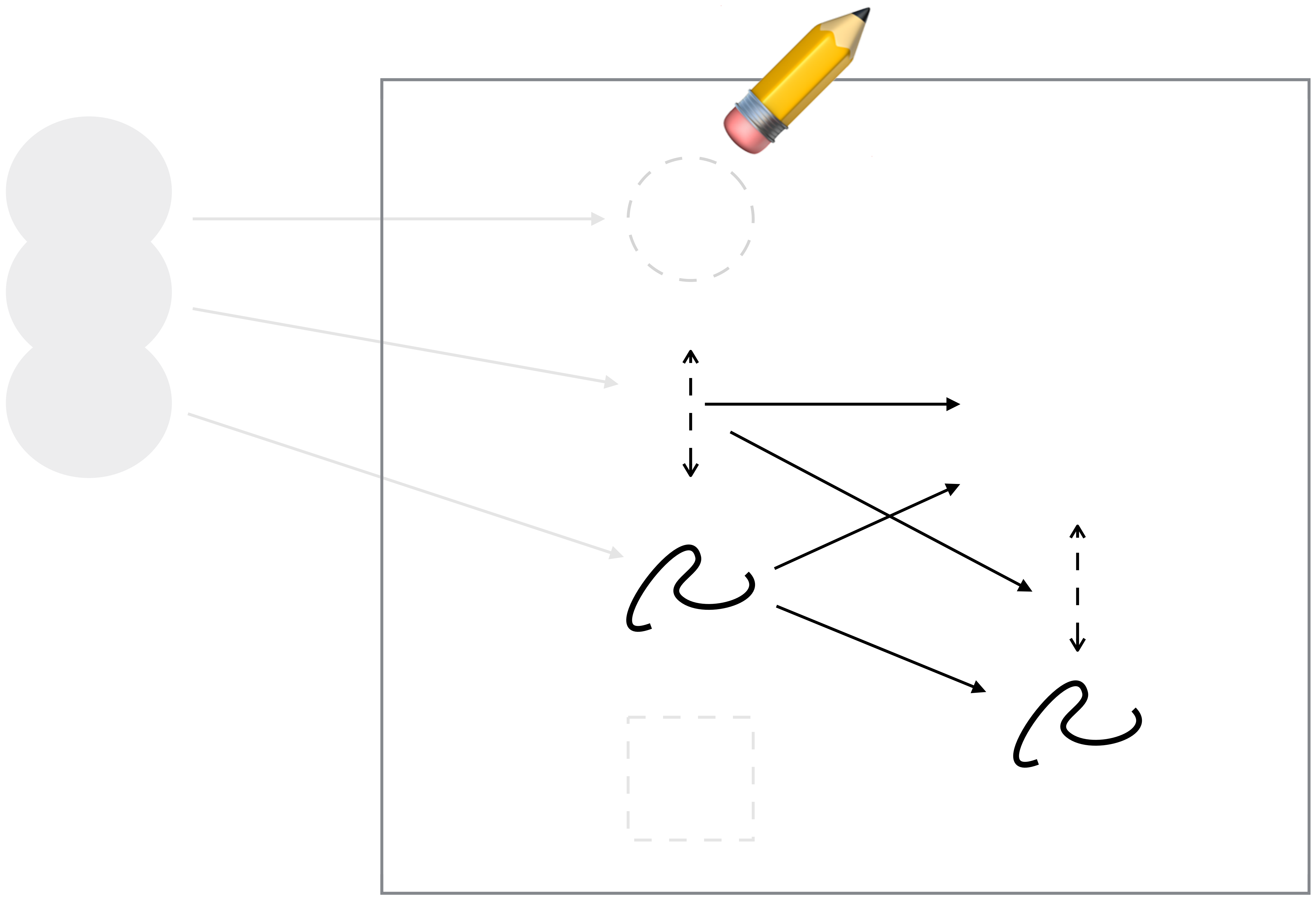


Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

Causality of Constituents

Composition across layers?



~~“dax”~~

Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

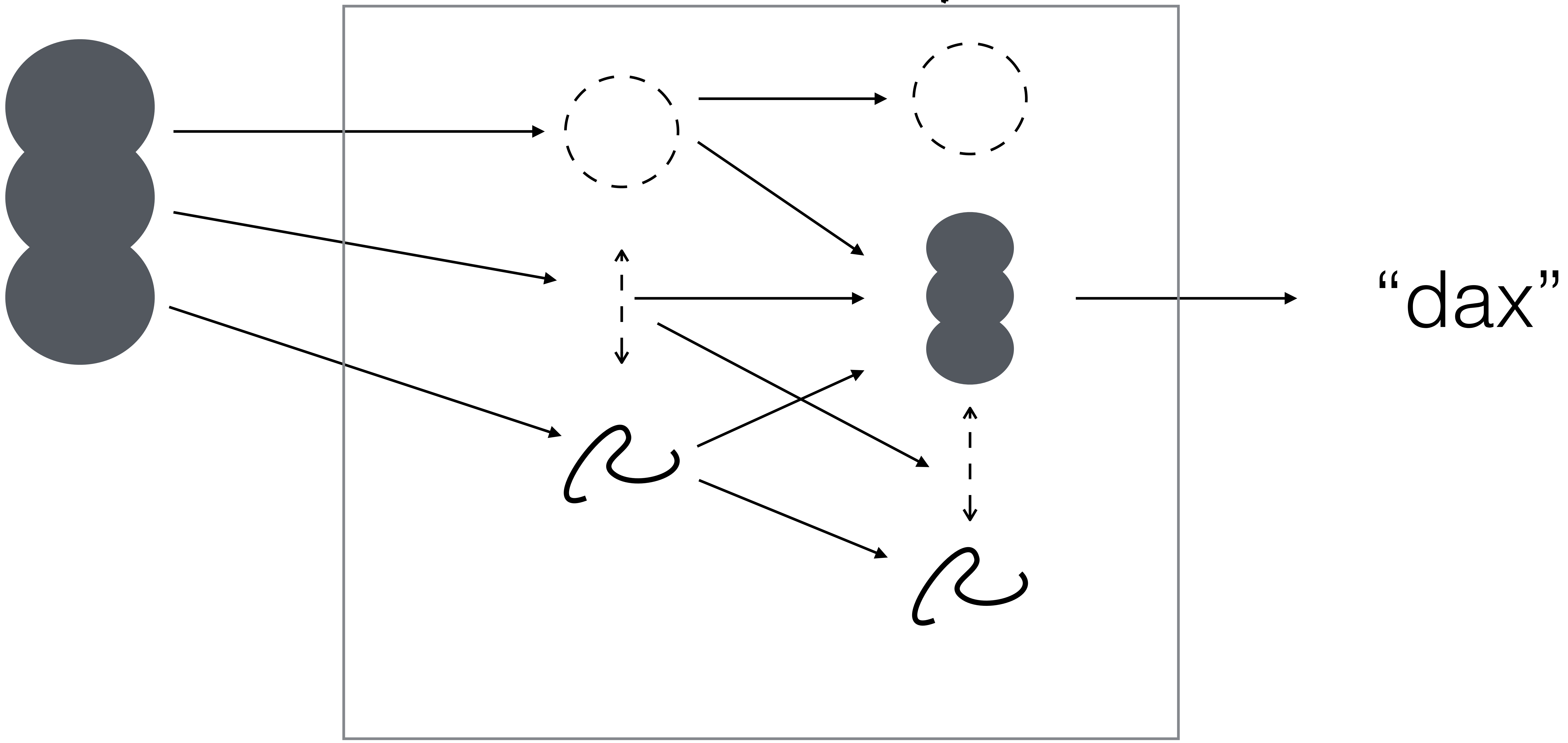
Causality of Constituents

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

Case Study

Causality of Constituents

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

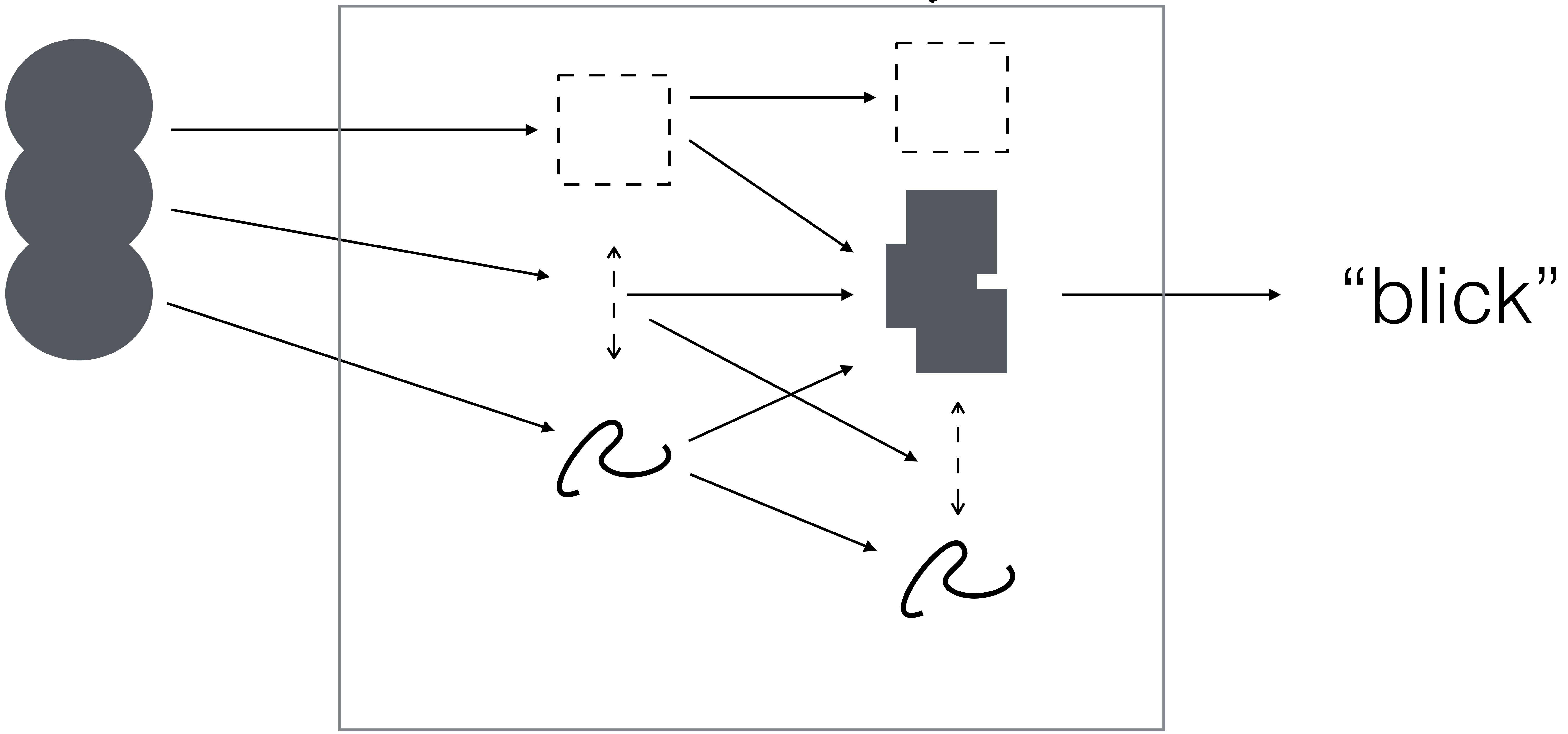


Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

Causality of Constituents

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



Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

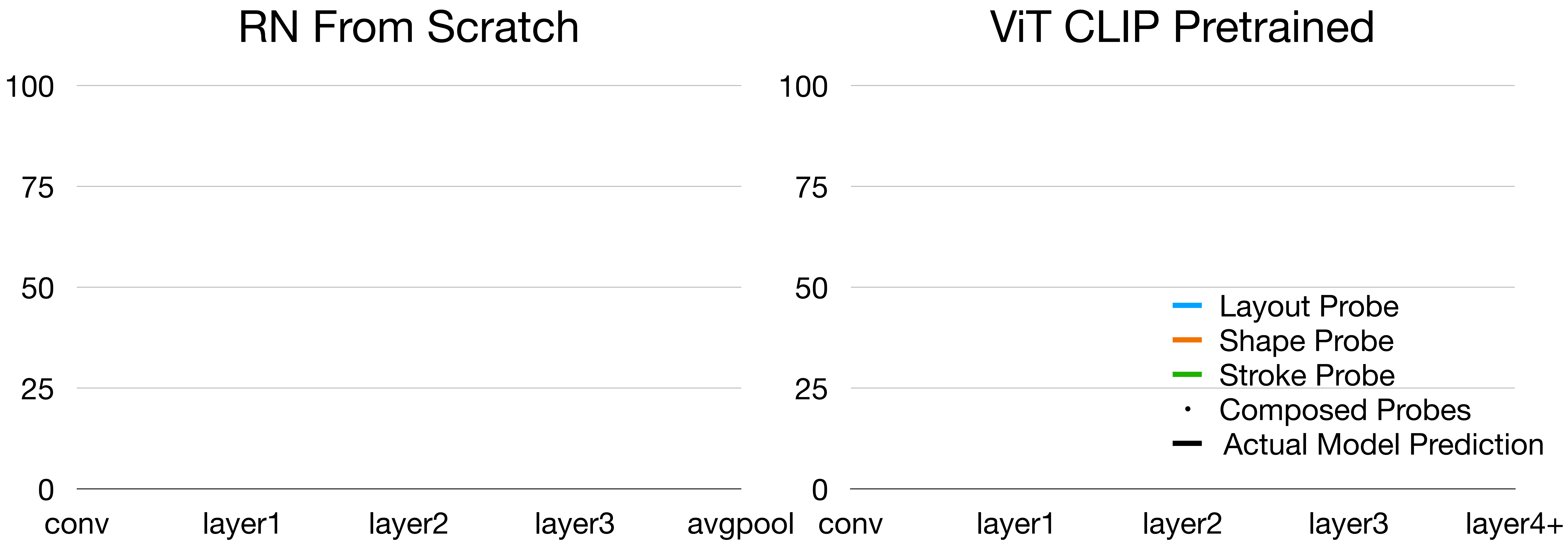
Case Study

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

Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

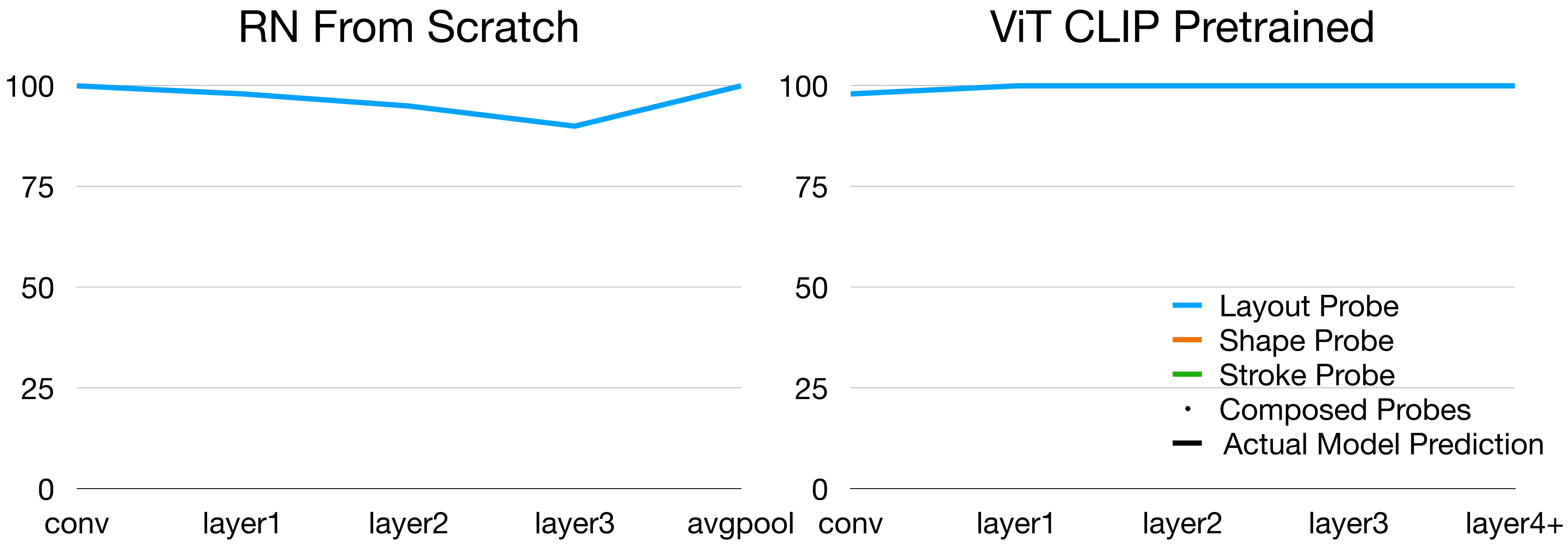
Can errors in the whole be explained by errors in the parts in aggregate?



Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

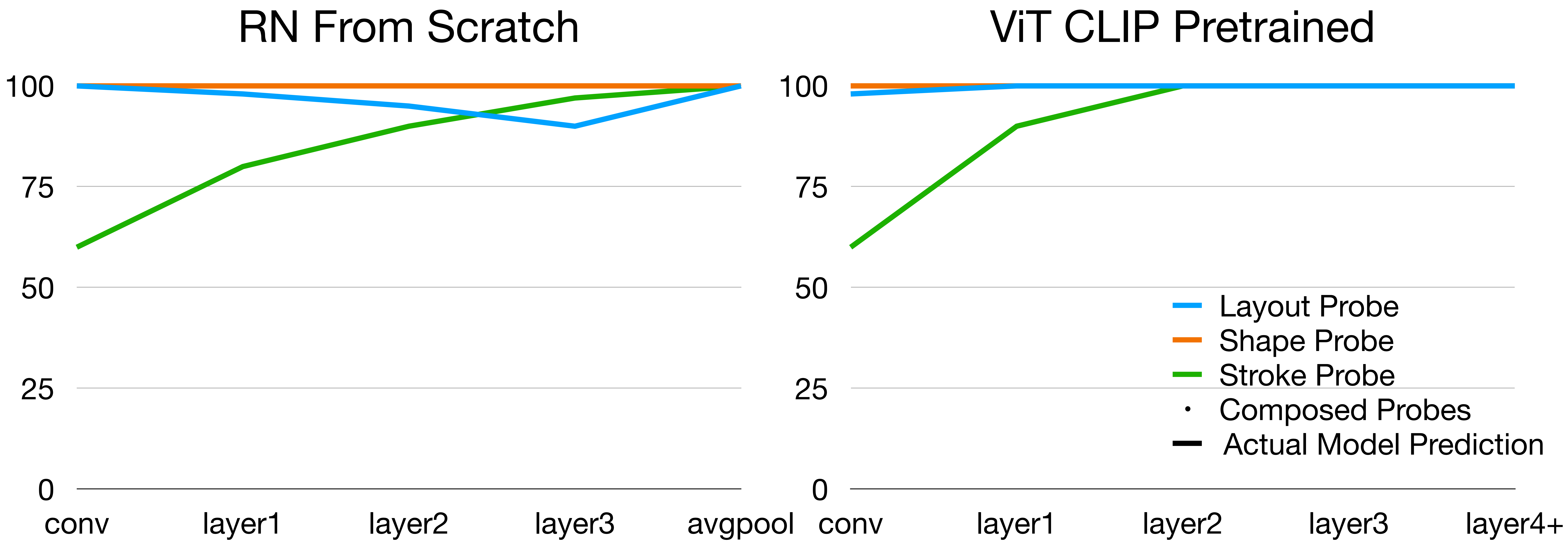
Can errors in the whole be explained by errors in the parts in aggregate?



Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

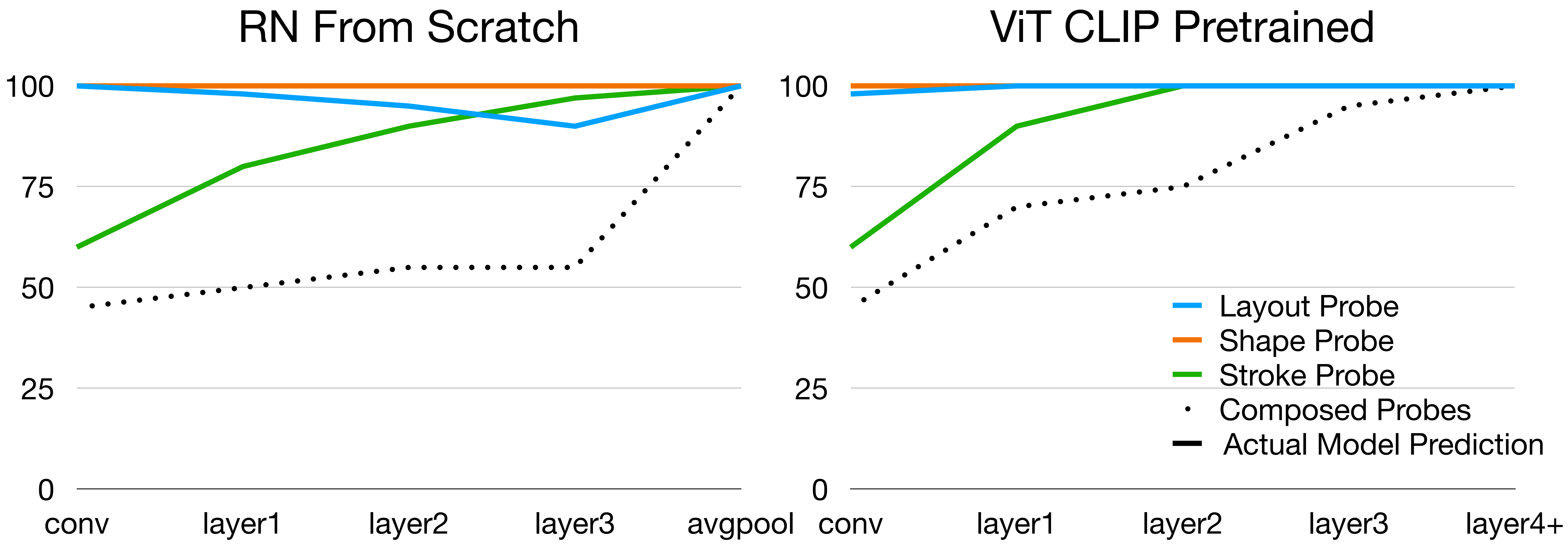
Can errors in the whole be explained by errors in the parts in aggregate?



Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

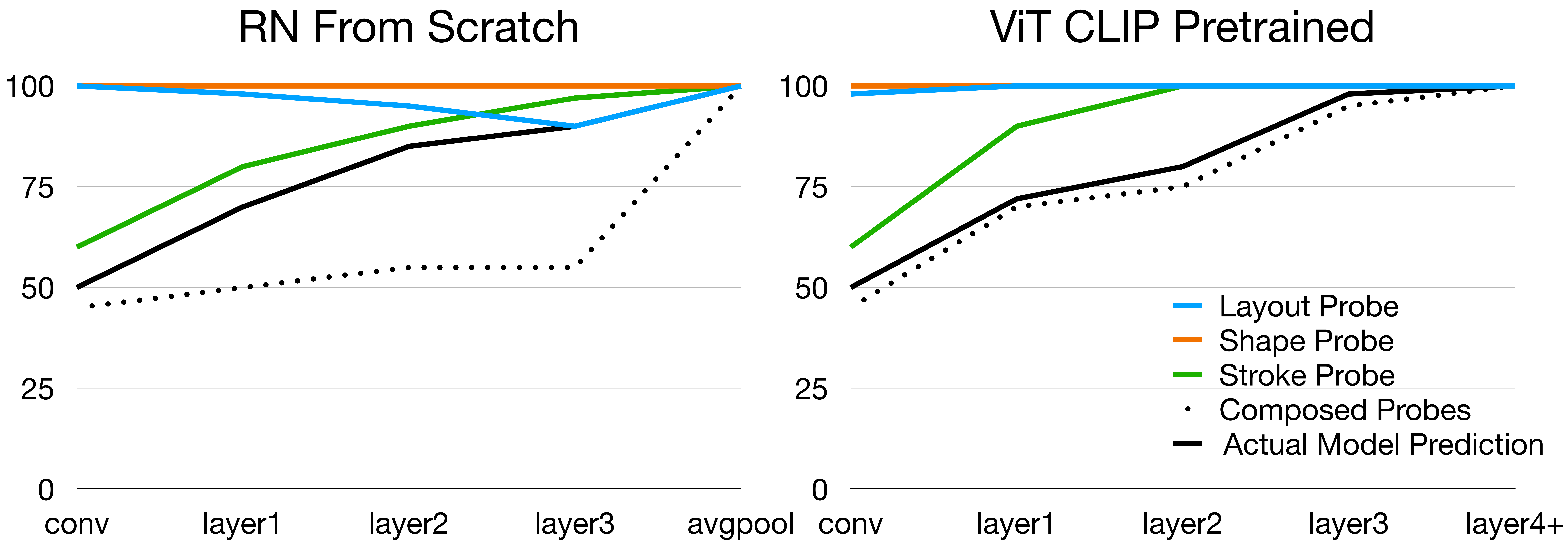
Can errors in the whole be explained by errors in the parts in aggregate?



Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

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



Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

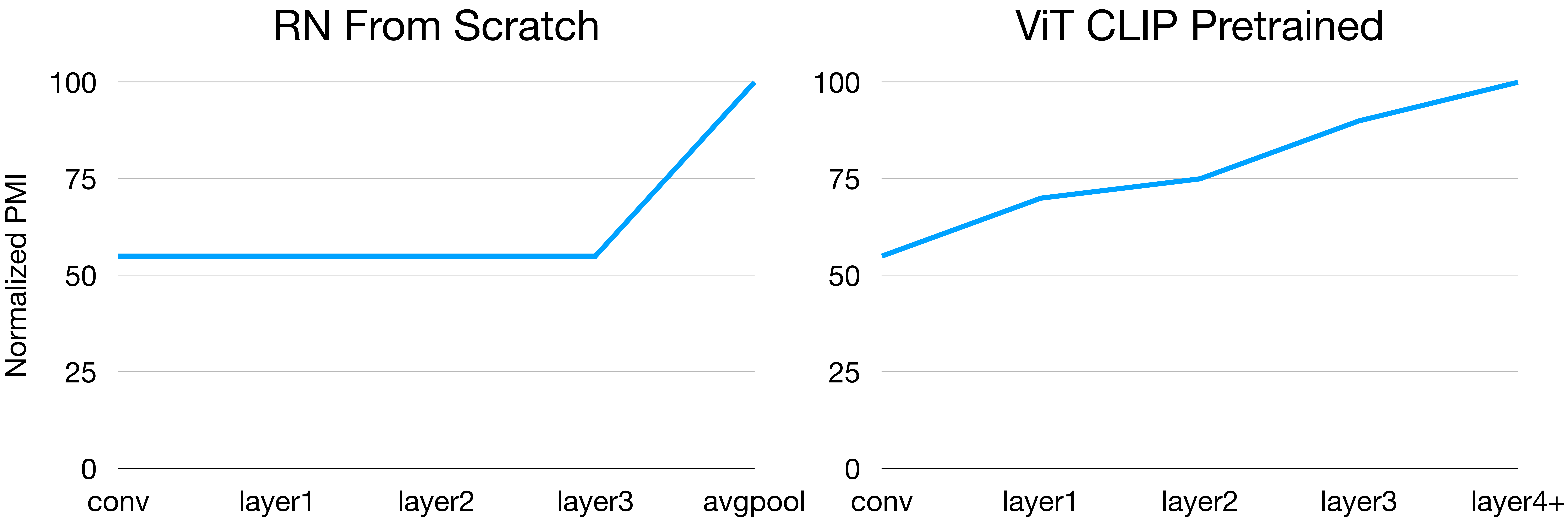
Case Study

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

Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

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



Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. TACL 2022 (to appear).

Case Study

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 build 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

Discussion

- NNs' representations are "**points in space**", but these points arguably can be understood as **structured representations** consisting of **reusable constituent parts**
- Determining the exact form of these representations take requires using **empirical measures other than behavior**
- There is serious **methodological development** required to build and vet these new empirical measures, but we have already begun and its within reach
- Whether these models meet the critiera of "compositional" requires serious **theoretical development**. I don't think the earlier debates anticipated models quite like this, and thus there is still work to do to refine definitions in order to know whether the current models are capable of giving us what we want.

Thank you!


Charles Lovering


Dylan Ebert


Jack Murello


Aaron Traylor


Sydney Zink


Albert Webson


Roma Patel


Jason Wei


Rohan Jha


Qinan Yu


Alyssa Loo

Conceptal Abstractions in NNs

Evaluating Large Language Models

Grounded Concept Learning