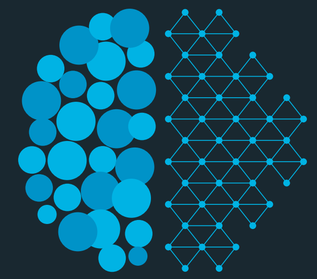
Institut
des algorithmes
d'apprentissage
de Montréal



MILA

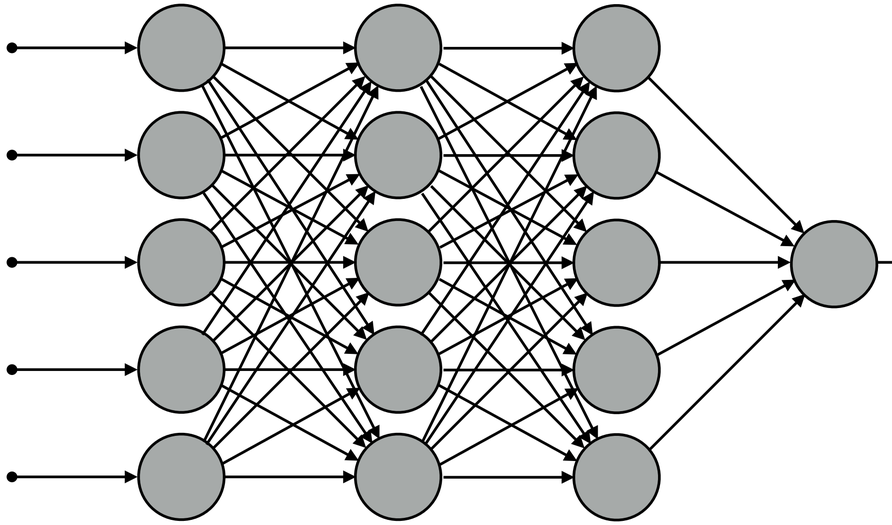
Systematic Generalization

What is it and how do we achieve it?

Aaron Courville
Mila & Université de Montréal

EEML
July 14th, 2021

Machine learning pipeline



Goldfish

Neural Network

clarifai ABOUT PRICING DEVELO

cow milk agriculture farm cattle livestock dairy
beef hayfield field grass mammal pasture calf
farmland rural animal pastoral bull grassland

cow beef agriculture cattle milk pasture mammal
livestock farmland grass farm hayfield rural herd
dairy pastoral grassland field calf bull

From Bernhard Schölkopf's slides (2017) - who got it from Pietro Perona (2017)

clarifai ABOUT PRICING DEVELO

cow milk agriculture farm cattle livestock dairy
beef hayfield field grass mammal pasture calf
farmland rural animal pastoral bull grassland

cow beef agriculture cattle milk pasture mammal
livestock farmland grass farm hayfield rural herd
dairy pastoral grassland field calf bull

From Bernhard Schölkopf's slides (2017) - who got it from Pietro Perona (2017)

clarifai ABOUT PRICING DEVELOPE

beach sand travel no person water sea seashore
summer sky outdoors ocean nature

water no person beach seashore sea sand mammal
outdoors travel ocean surf sky

From Bernhard Schölkopf's slides (2017) - who got it from Pietro Perona (2017)

clarifai ABOUT PRICING DEVELOPE

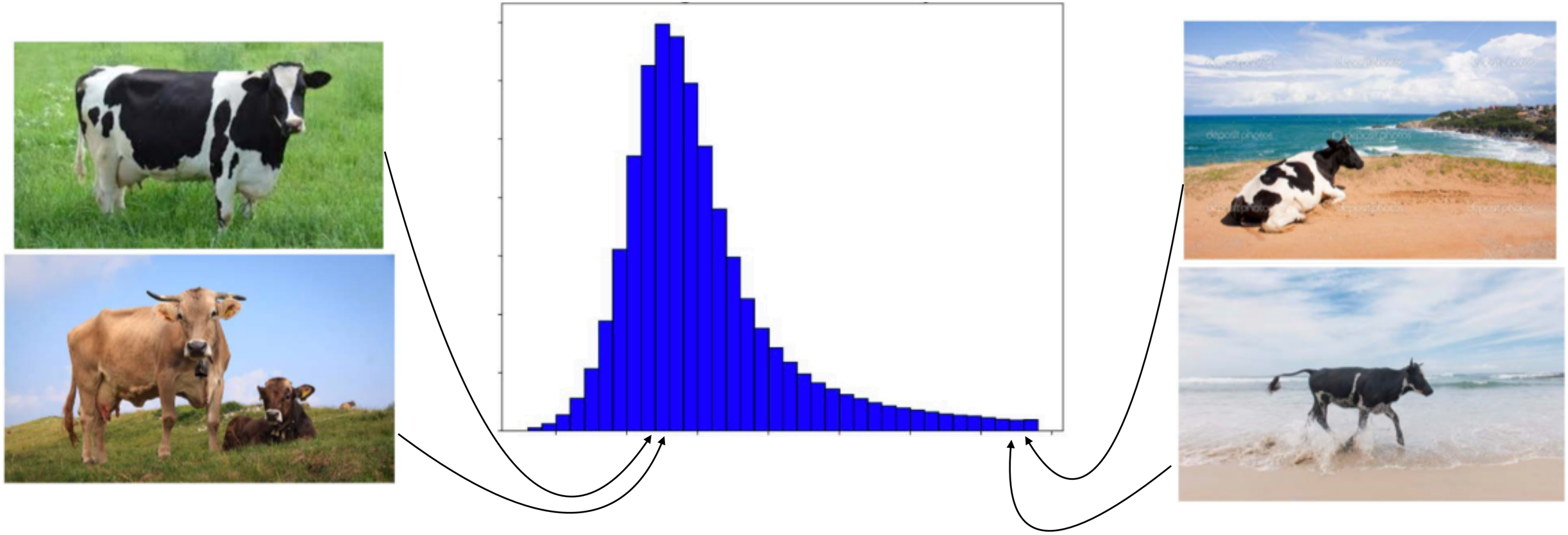
beach sand travel no person water sea seashore
summer sky outdoors ocean nature

water no person beach seashore sea sand mammal
outdoors travel ocean surf sky

From Bernhard Schölkopf's slides (2017) - who got it from Pietro Perona (2017)

How generalization fails

- Generalization failures often result from **out-of-distribution** test settings.



- Learner becomes distracted by correlated but spurious features.

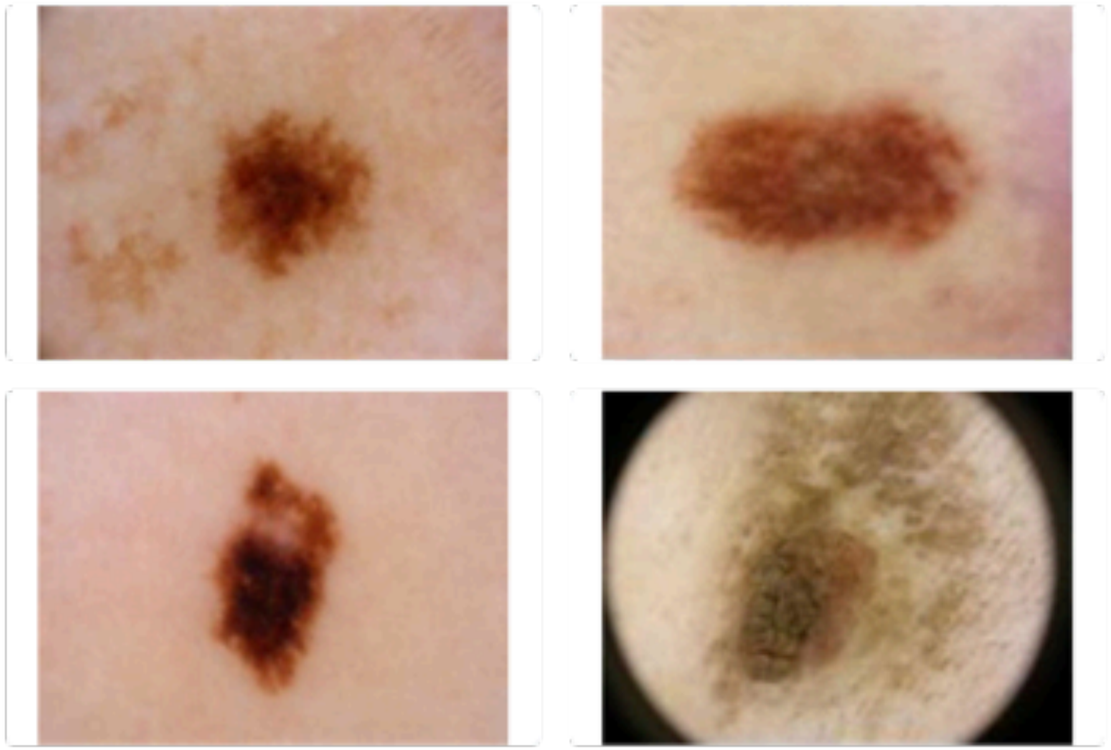
Data Bias?

- Biased training datasets are often blamed for the failure to generalize at test time (or deployment).
  - Example: **Arjovsky et al. (2019)** *Invariant Risk Minimization*

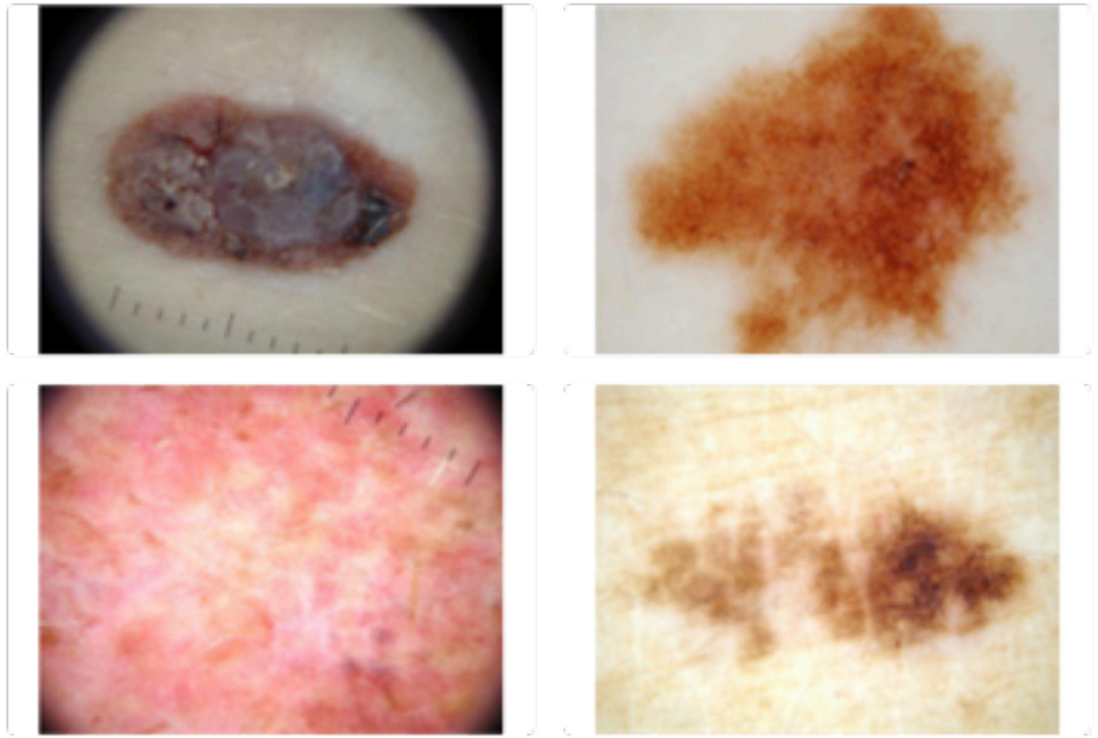
- Sometimes the data is truly biased…. But sometimes not.

Data Bias

- Ruler marks that show up in malignant tumours in the training set are learned, by a CNN, to be important features predicting malignancy



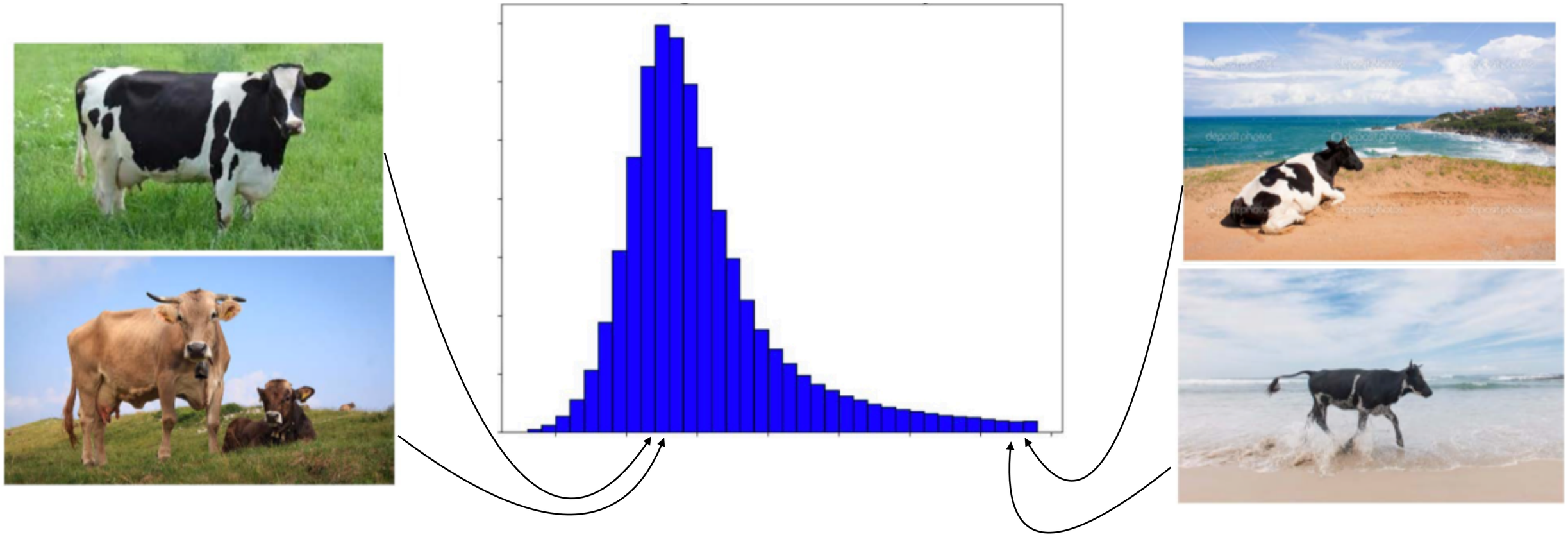
benign lesions



malignant cancers

Images from Nicole Wheeler (2019) article in Towards Data Science, originally sourced from www.isic-archive.com

How generalization fails without data bias



- Here generalization failure can result from **out-of-distribution** test setting w.r.t. the **empirical training distribution**.

Dealing with OoD Generalization

- Out-of-distribution generalization emerges because training and test data are drawn from different settings, domains or **environments**.

- **Broad Strategy**: embrace the non-iid nature of the data at training time and learn to be invariant to changes across these environments.

- Three approaches to this (Wang et al, 2021 - https://arxiv.org/abs/2103.03097):

  1. **Data manipulation** (eg. data-augmentation, domain randomization)
  2. **Representation learning** (eg. adversarial domain adapt., invariant risk minimization)
  3. **Learning strategies** (eg. meta-learning)

OoD Benchmark: DomainBed

Dataset	Domains					
Colored MNIST	+90%	+80%	-90%			
						
	(degree of correlation between color and label)					
Rotated MNIST	0°	15°	30°	45°	60°	75°
						
VLCS	Caltech101	LabelMe	SUN09	VOC2007		
						
PACS	Art	Cartoon	Photo	Sketch		
						
Office-Home	Art	Clipart	Product	Photo		
						
Terra Incognita	L100	L38	L43	L46		
						
	(camera trap location)					
DomainNet	Clipart	Infographic	Painting	QuickDraw	Photo	Sketch
						

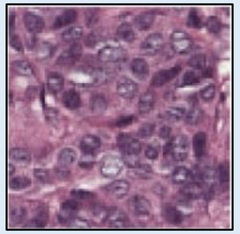
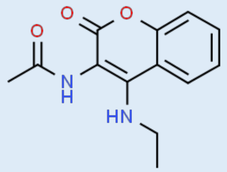
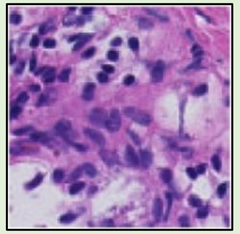
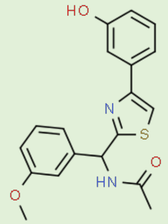
In search of lost domain generalization, Gulrajani and Lopez-Paz, ICLR 2021

https://github.com/facebookresearch/DomainBed

OoD Benchmark: WILDS

https://github.com/p-lambda/wilds

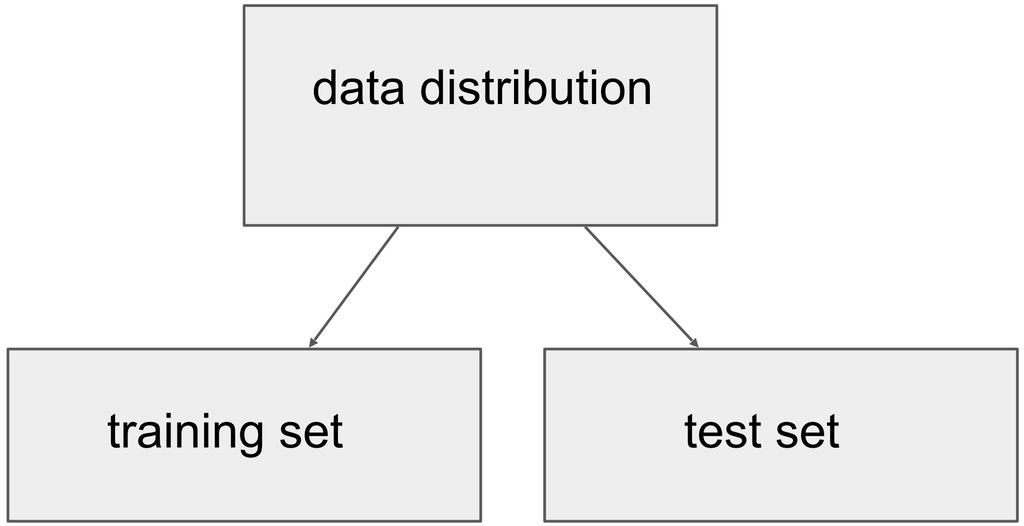
WILDS: A benchmark of in-the-wild distribution shifts, Koh and Sagawa et al., 2020

	Domain generalization			Subpopulation shift	Domain generalization + subpopulation shift			
Dataset	iWildCam	Camelyon17	OGB-MolPCBA	CivilComments	FMoW	PovertyMap	Amazon	Py150
Input (x)	camera trap photo	tissue slide	molecular graph	online comment	satellite image	satellite image	product review	code
Prediction (y)	animal species	tumor	bioassays	toxicity	land use	asset wealth	sentiment	autocomplete
Domain (d)	camera	hospital	scaffold	demographic	time, region	country, rural-urban	user	git repository
# domains	323	5	120,084	16	16 x 5	23 x 2	2,586	8,421
# examples	203,029	455,954	437,929	448,000	523,846	19,669	539,502	150,000
Train example				What do Black and LGBT people have to do with bicycle licensing?			Overall a solid package that has a good quality of construction for the price.	`import numpy as np … norm=np.___`
Test example				As a Christian, I will not be patronizing any of those businesses.			I *loved* my French press, it's so perfect and came with all this fun stuff!	`import subprocess as sp p=sp.Popen() stdout=p.___`
Adapted from	Beery et al. 2020	Bandi et al. 2018	Hu et al. 2020	Borkan et al. 2019	Christie et al. 2018	Yeh et al. 2020	Ni et al. 2019	Raychev et al. 2016

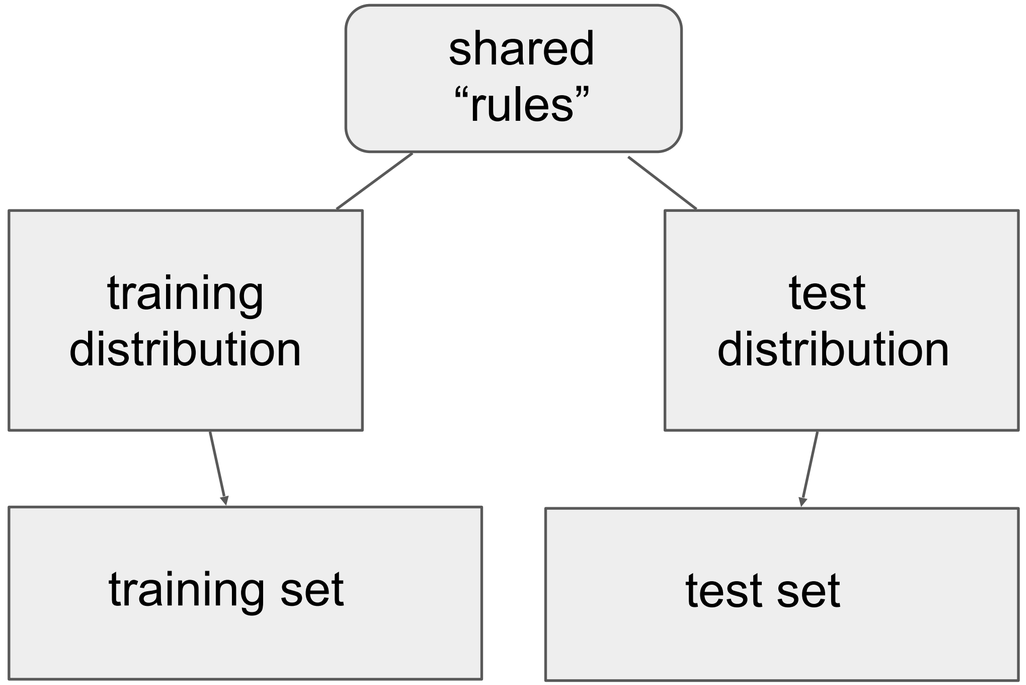
Systematic Generalization

- **Systematic Generalization**: Generalization to examples that may not be drawn from the same distribution as the training data, but that obey the same basic rules of production or underlying structure.

Generalization:



Systematic Generalization:



Systematic Generalization in Language

Lake and Baroni (2018): Generalization without Systematicity

- **Systematic Compositionality:** The capacity to understand and produce a potentially infinite number of novel combinations from known components.
  - Example: (from Lake and Baroni, 2018)
    Consider you've just learned meaning of a new verb *"to dax"*
    You understand the meaning of *"dax twice and then dax again"*

- Modern neural networks display impressive generalization.
  Question: **Do they demonstrate systematicity?**
  Answer (L&B2018): No they don't.

Systematic Generalization in Language

Lake and Baroni (2018): Generalization without Systematicity

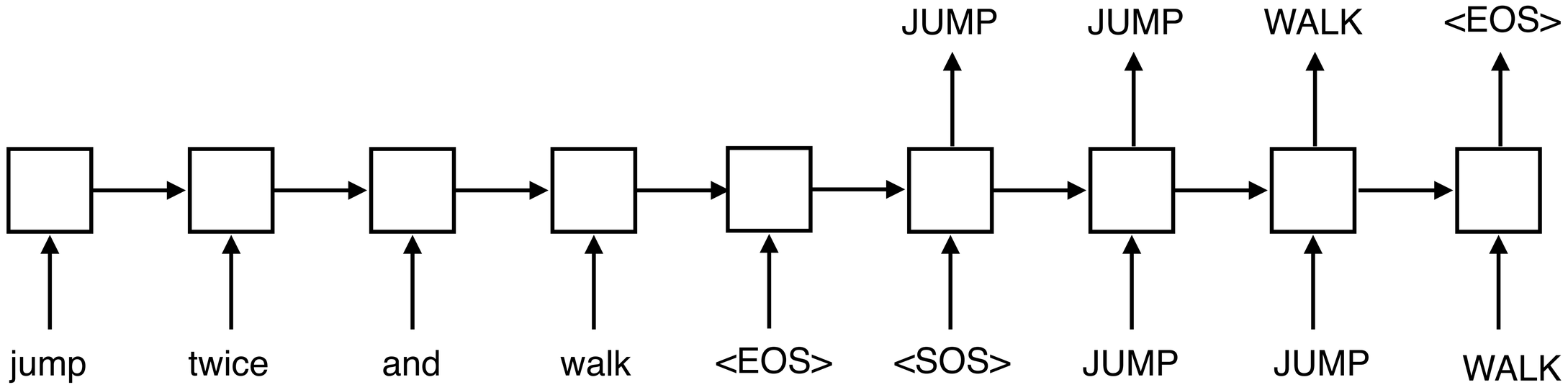
- Develop the SCAN dataset and task.

jump	⇒	JUMP
jump left	⇒	LTURN JUMP
jump around right	⇒	RTURN JUMP RTURN JUMP RTURN JUMP RTURN JUMP
turn left twice	⇒	LTURN LTURN
jump thrice	⇒	JUMP JUMP JUMP
jump opposite left and walk thrice	⇒	LTURN LTURN JUMP WALK WALK WALK
jump opposite left after walk around left	⇒	LTURN WALK LTURN WALK LTURN WALK LTURN WALK LTURN LTURN JUMP

Systematic Generalization in Language

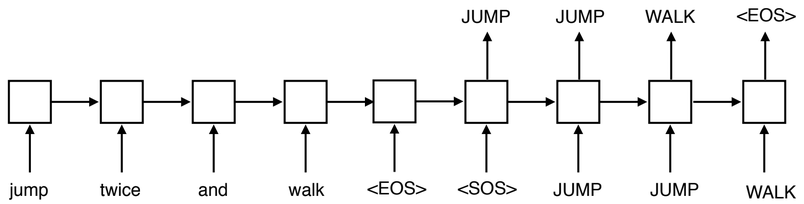
Lake and Baroni (2018): Generalization without Systematicity

- Develop the SCAN dataset and task.
- Trained Seq2Seq models (RNNs, actually LSTM) on this “translation” task.



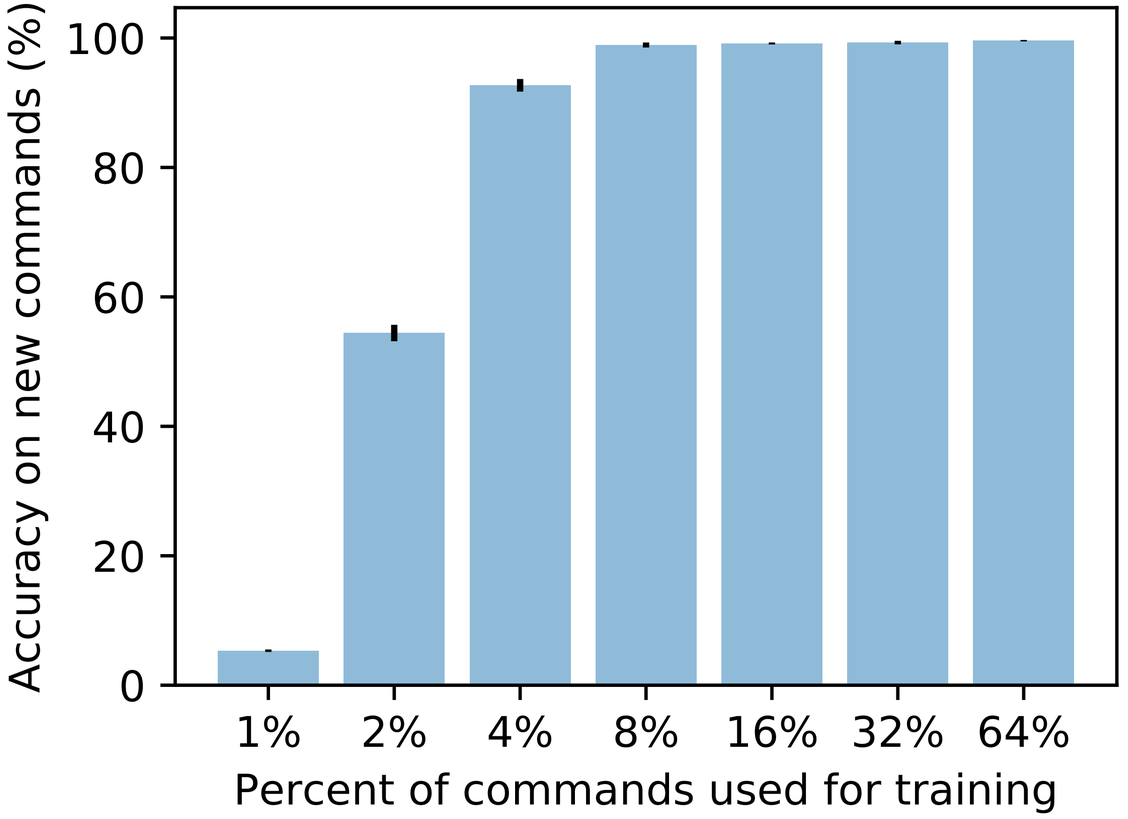
Systematic Generalization in Language

Lake and Baroni (2018): Generalization without Systematicity



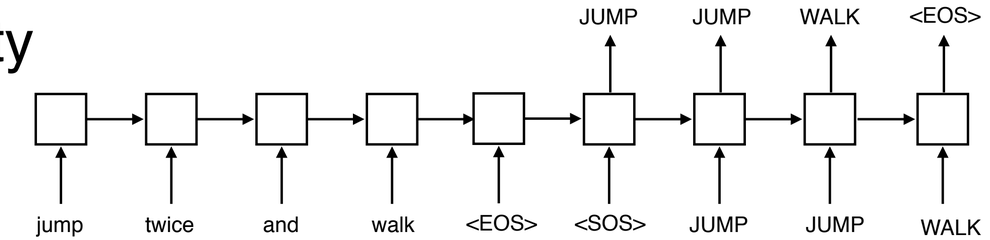
Baseline Experiment (i.e. Standard Generalization):

- Randomly sample a training set of commands, test on the rest.



Systematic Generalization in Language

Lake and Baroni (2018): Generalization without Systematicity



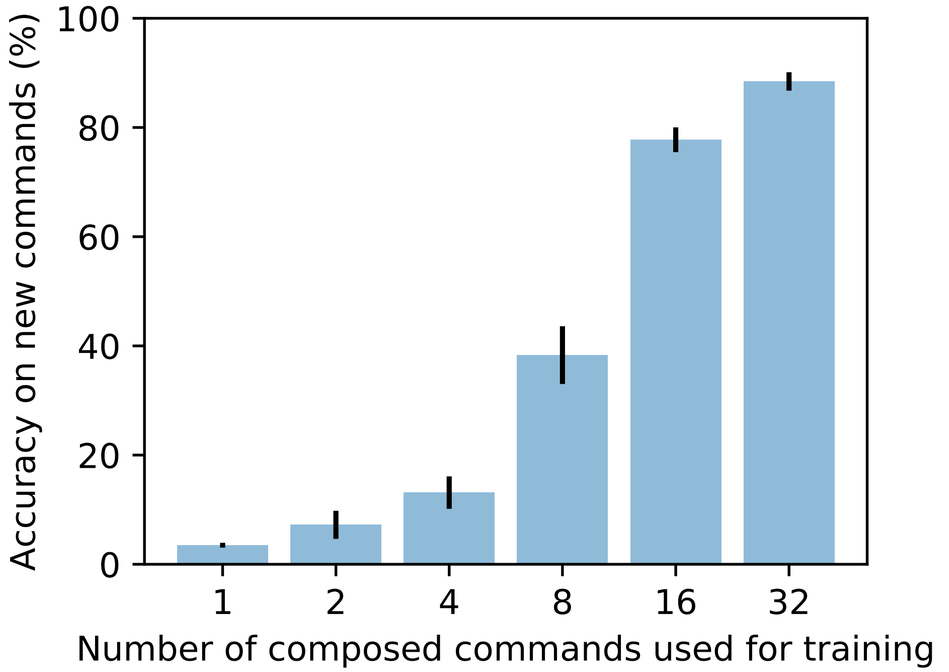
Training phase (***add jump*** split):

- For one action (e.g., “jump”): trained on only primitive commands and a few composed commands.
- For all other actions: trained on all primitive and composed commands for all other actions (e.g., “run”, “run twice”, “walk”, “walk opposite left and run twice”, etc.).

Test Phase (***add jump*** split):

- Execute all other composed commands for the primitive action (e.g., “jump twice”, “jump opposite left and run twice”, etc.).

Compositionality: if you know the meaning of “run”, “jump” and “run twice”, you should also understand what “jump twice” means.



Promoting Systematic Generalization I

1. Increase data diversity / Data augmentation:

   - Hill et al. (2019) *Environmental Drivers of Systematicity and Generalization in a Situated Agent.*
   - Andreas (2020) Good Enough Compositional Data Augmentation.

2. Inductive bias / Strong priors:

   - Tan, Shen et al. (2020) *Recursive Top-Down Production for Sentence Generation with Latent Trees*.
   - Lui et al. (2020) *Compositional Generalization by Learning Analytical Expressions.*

3. Meta learning:

   - Lake (2019) *Compositional generalization through meta sequence-to-sequence learning.*
   - Conklin et al (2021) *Meta-learning to Compositionally Generalize.*

Promoting Systematic Generalization II

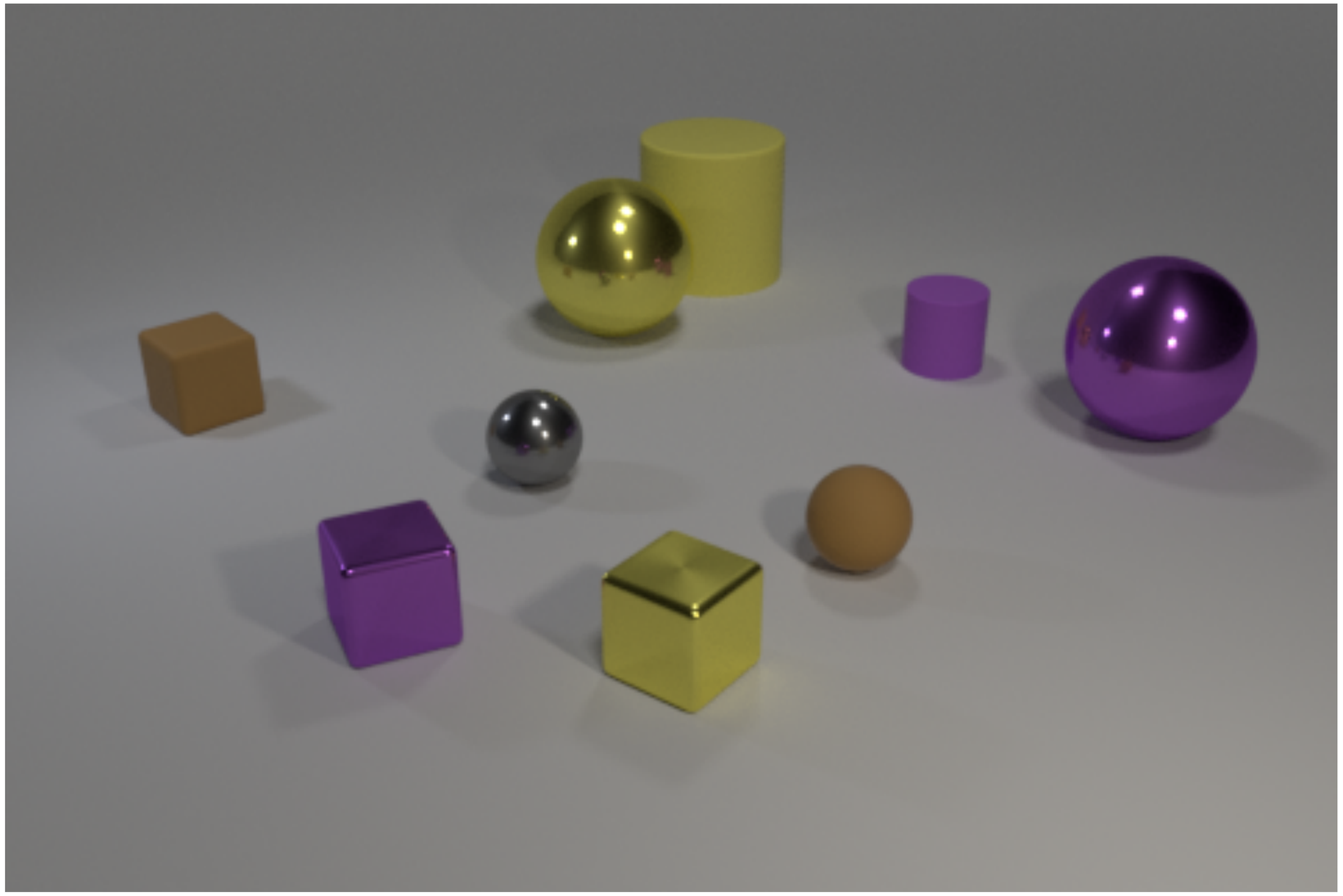
4. Modularity

   - **Neural Modular Networks (NMN)** (Andreas et al. 2016) dynamically compose novel architectures to perform inference / prediction.

5. **Iterated Learning** (Kirby et al. papers starting from around 2000)

   - Language emergence theory from cognitive science shows promise for learning compositional representations.

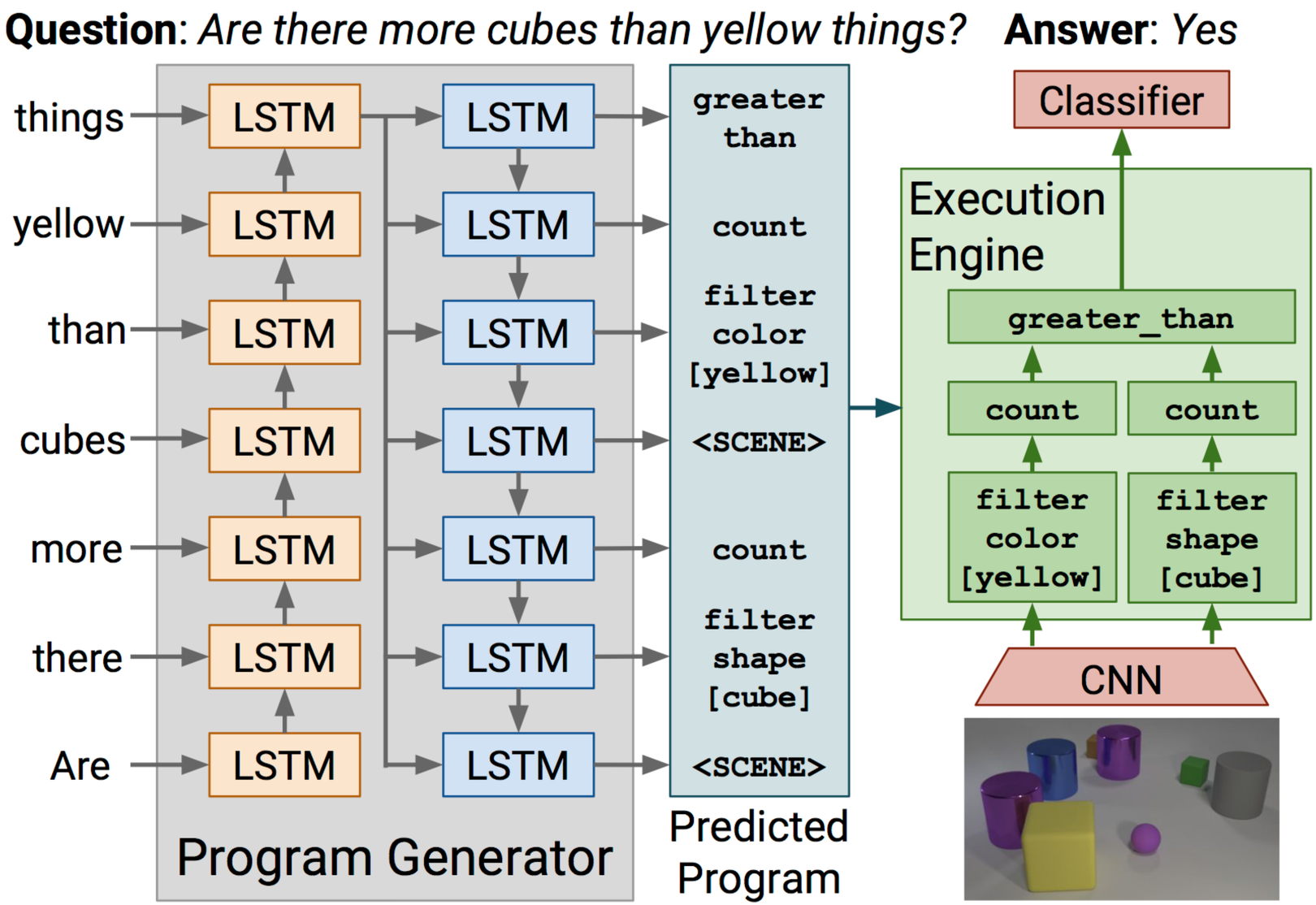
CLEVR Visual Reasoning Dataset



Q: What number of cylinders are small purple things or yellow rubber things?

A: 2

Neural Module Networks



Johnson et al.
Inferring and Executing Programs for Visual Reasoning.
ICCV 2017.

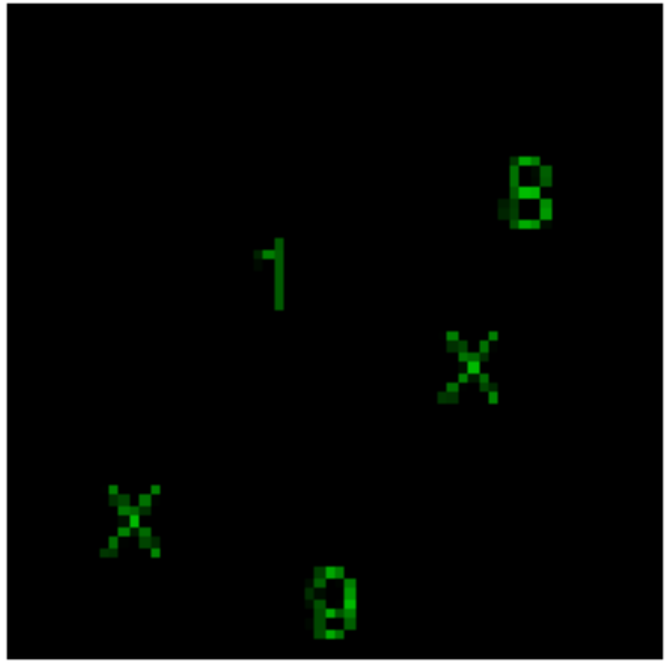
SQOOP: Systematic Generalization

Dzmitry Bahdanau, Shikhar Murty, Michael Noukhovitch,Thien Huu Nguyen and Harm de Vries, Aaron Courville (ICLR 2019)

- 64x64 RGB images of letters and digits (36 unique elements) with True/False (Yes / No) spatial relation question $q = X\ R\ Y$ (4 spatial relations: up, down, left, right).
- Distractor objects: $X' \neq X$ and $Y' \neq Y$ such that $X'\ R\ Y$ and $X\ R\ Y'$ are both true.

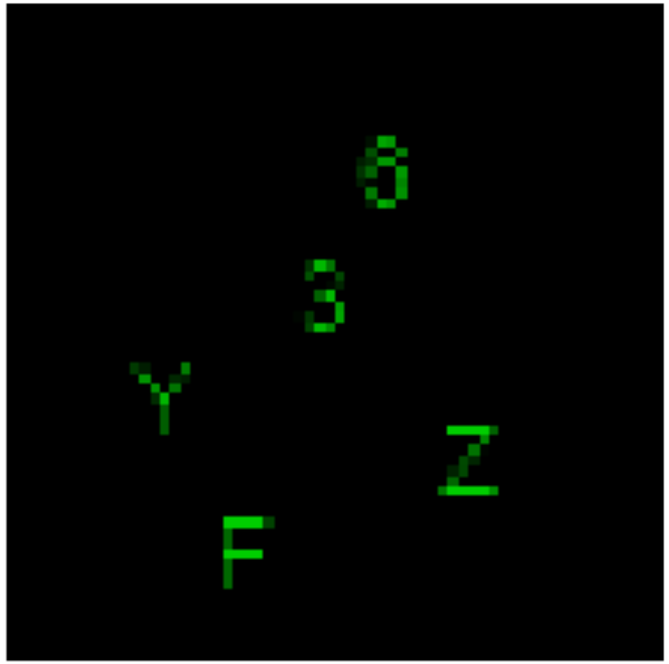
Is there a 9 right of a X?

True



Is there a 3 right of a 6?

False



SQOOP: Systematic Generalization

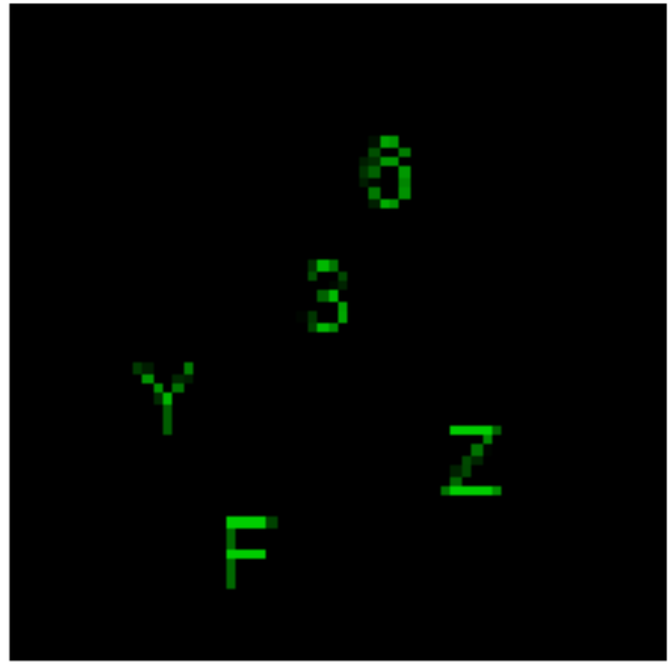
- 64x64 RGB images of letters and digits (36 unique elements) with True/False (Yes / No) spatial relation question $q = X\ R\ Y$.
- Distractor objects: $X' \neq X$ and $Y' \neq Y$ such that $X'\ R\ Y$ and $X\ R\ Y'$ are both true.

9 right_of X ?

True

3 right_of 6 ?

False



SQOOP: Systematic Generalization

Goal: evaluate generalization to all possible 36x36 possible object pairs.

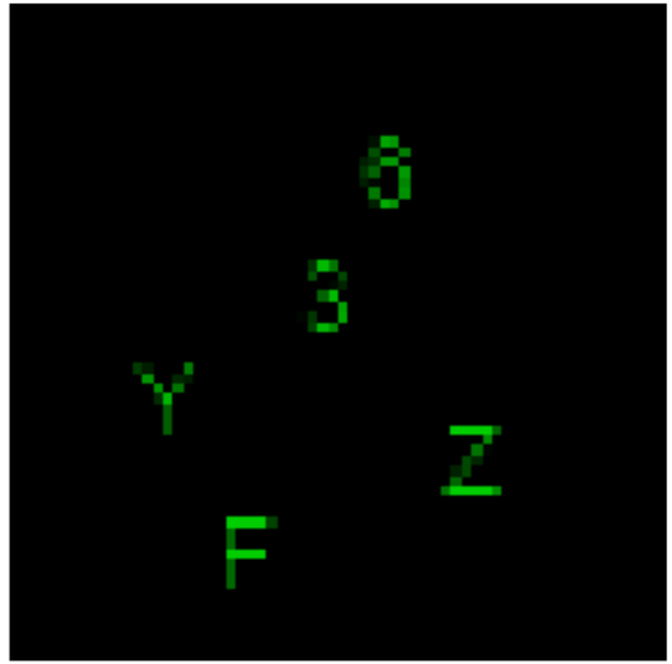
- Training set: 36 x 4 x *k* — *k* is the number of *rhs* objects for every *lhs*.
- Test set: 36 x 4 x (36 - *k*) — all pairs not in the training set.
- 1M training images — to avoid overfitting to training images.

9 right_of X ?

True

3 right_of 6 ?

False



Modeling Pardigm

Visual Input

CNN

Question

W left_of J?

Model:
FiLM / MAC / NMN

NN
Classifier

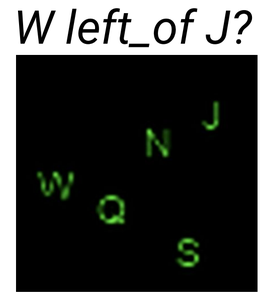
True
or
False

h_{x}

$h_{\mathrm{q\,x}}$

$$h_{\mathrm{x\,q}} = \mathrm{model}(h_{\mathrm{x}}, q)$$

Neural Module Networks



- Constructs a question-specific network by composing neural modules.
  - Need to resolve (i) the module types and number and (ii) their layout (computation graph)
- Neural module $f(\theta,\gamma,h_l,h_r)$ and question $q = (q_1,q_2,q_3)$
- Use 3 modules of a single binary module type.

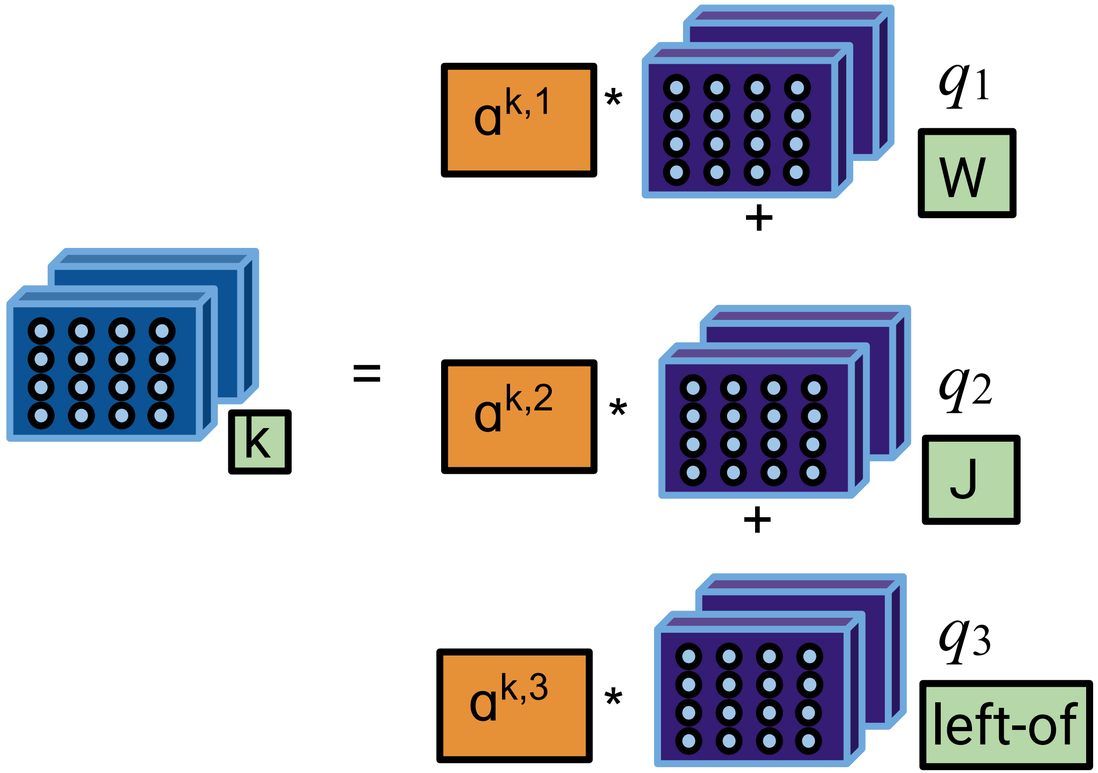
question attention
(assume 1-hot vectors)

$$\gamma_k = \sum_{i=1}^{s} \alpha^{k,i} e(q_i)$$

$$h_k = f(\theta, \gamma_k, \sum_{j=-1}^{k-1} \tau_0^{k,j} h_j, \sum_{j=-1}^{k-1} \tau_1^{k,j} h_j)$$

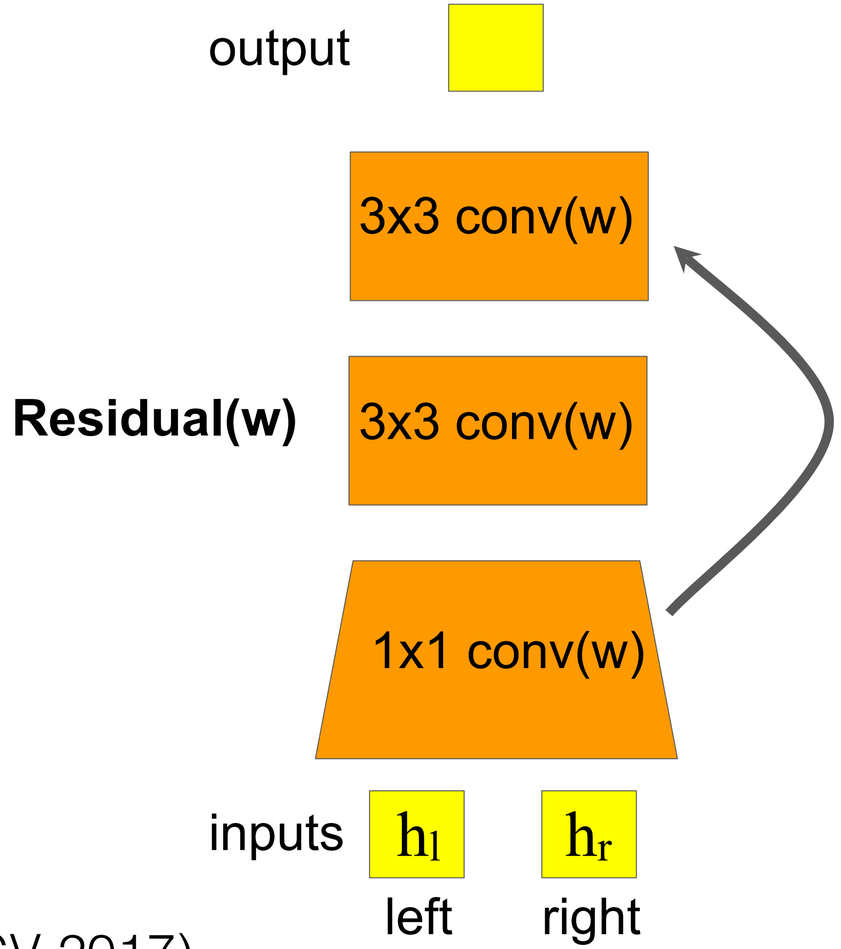
$$h_{qx} = h_M$$

layout parameters

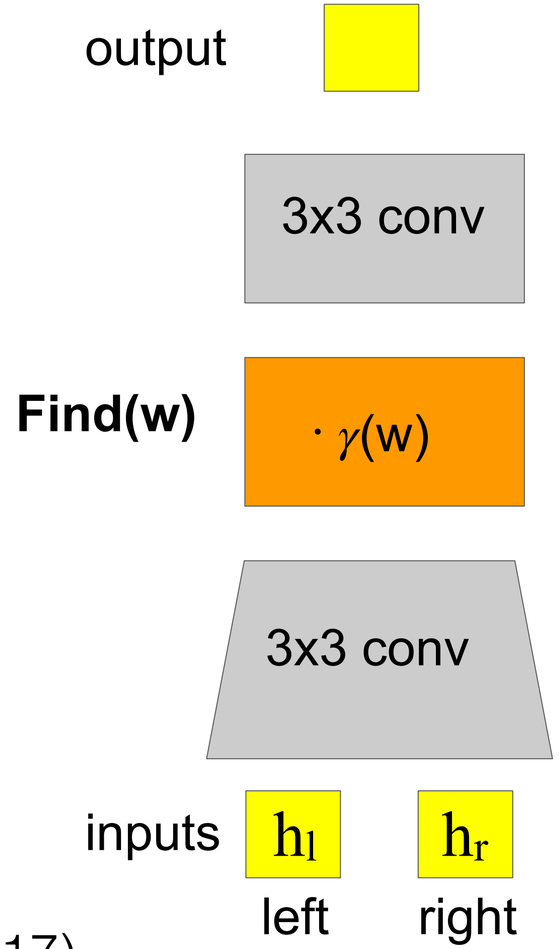


NMN Modules

- We explore two varieties of NMN modules $h_k = f(\theta, \gamma, h_l, h_r)$:



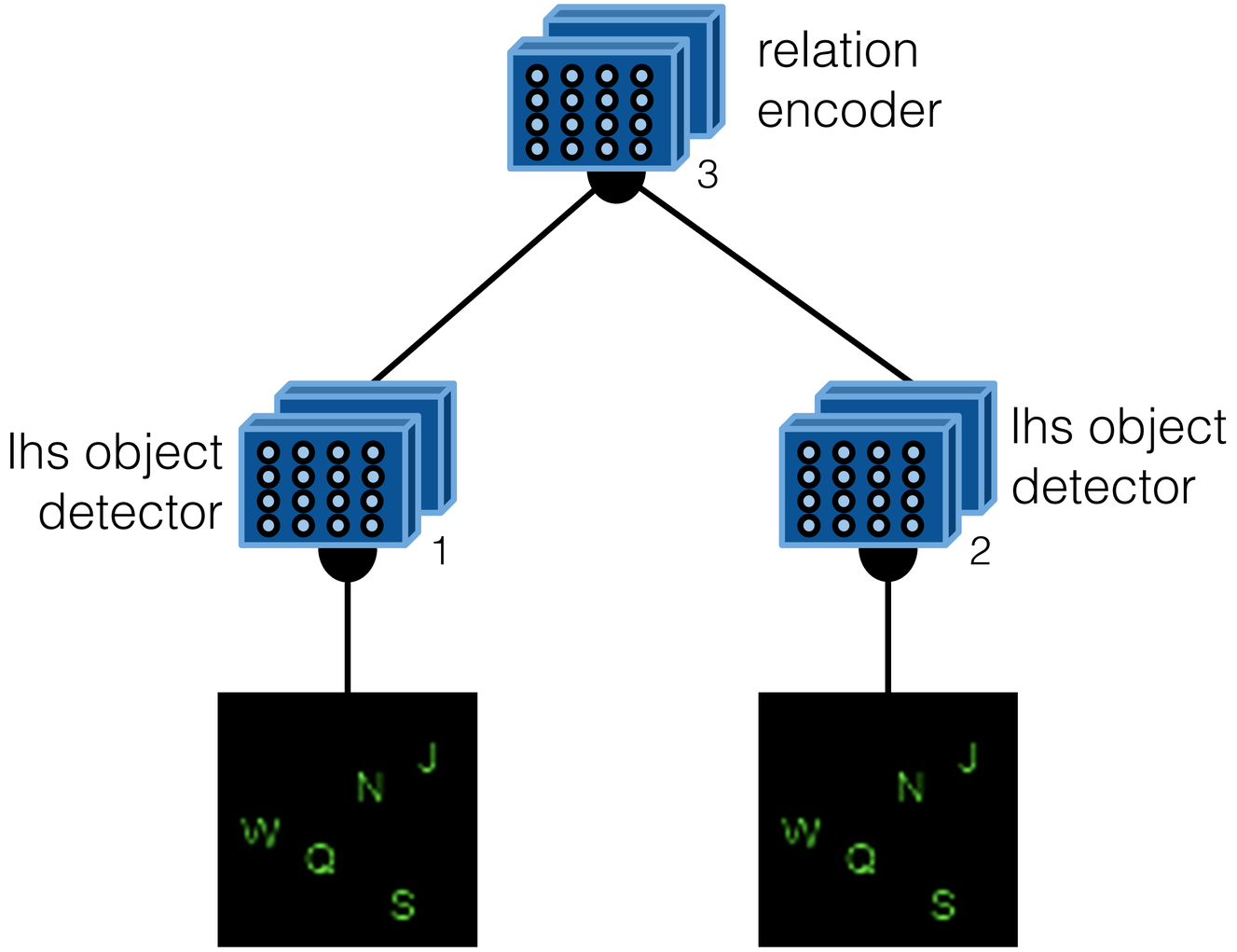
Johnson et al. (ICCV 2017)
Inferring and Executing Programs for Visual Reasoning.



Hu et al. (ICCV 2017)
Learning to Reason: End-to-End Module Networks for Visual Question Answering.

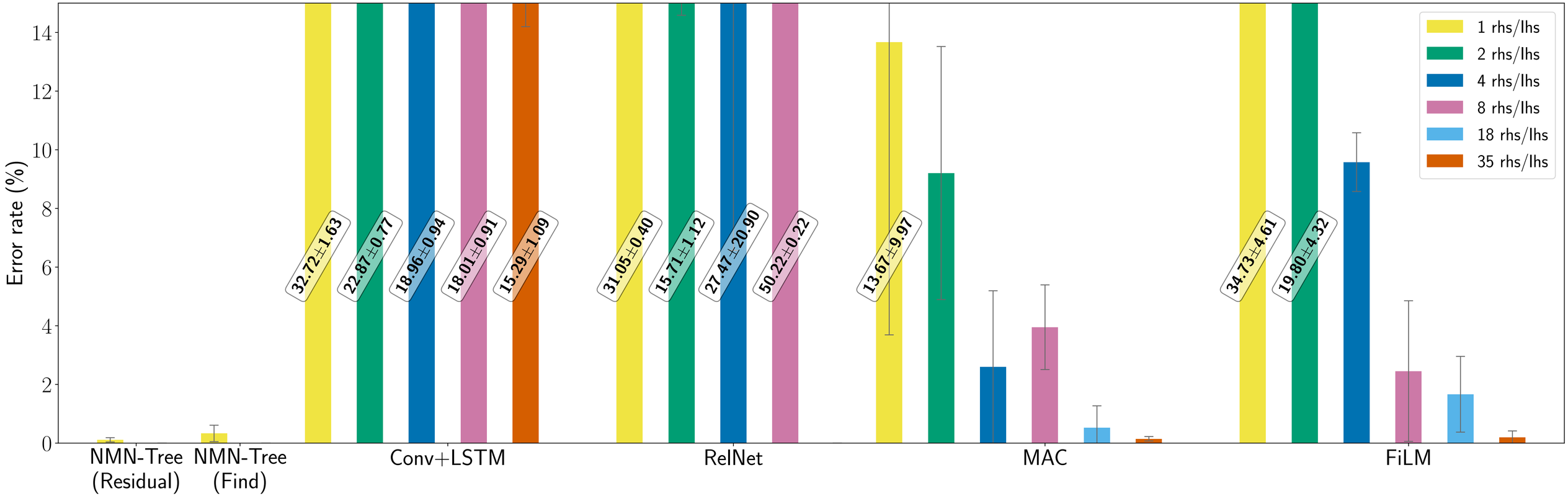
NMN layout: For SQOOP, tree-structure is optimal

W left_of J?

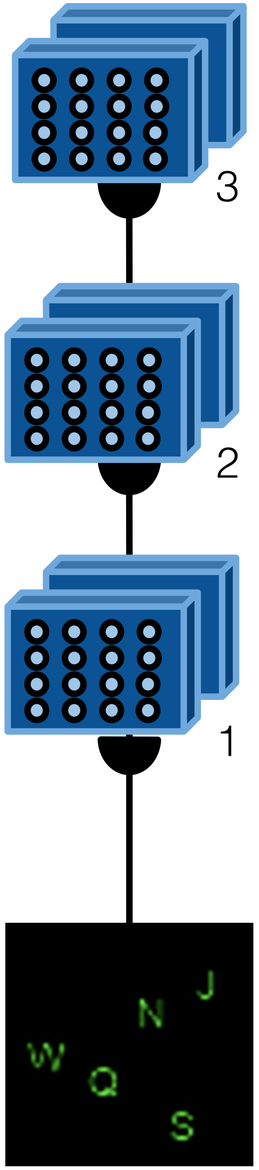


Experiments I: which models generalize better?

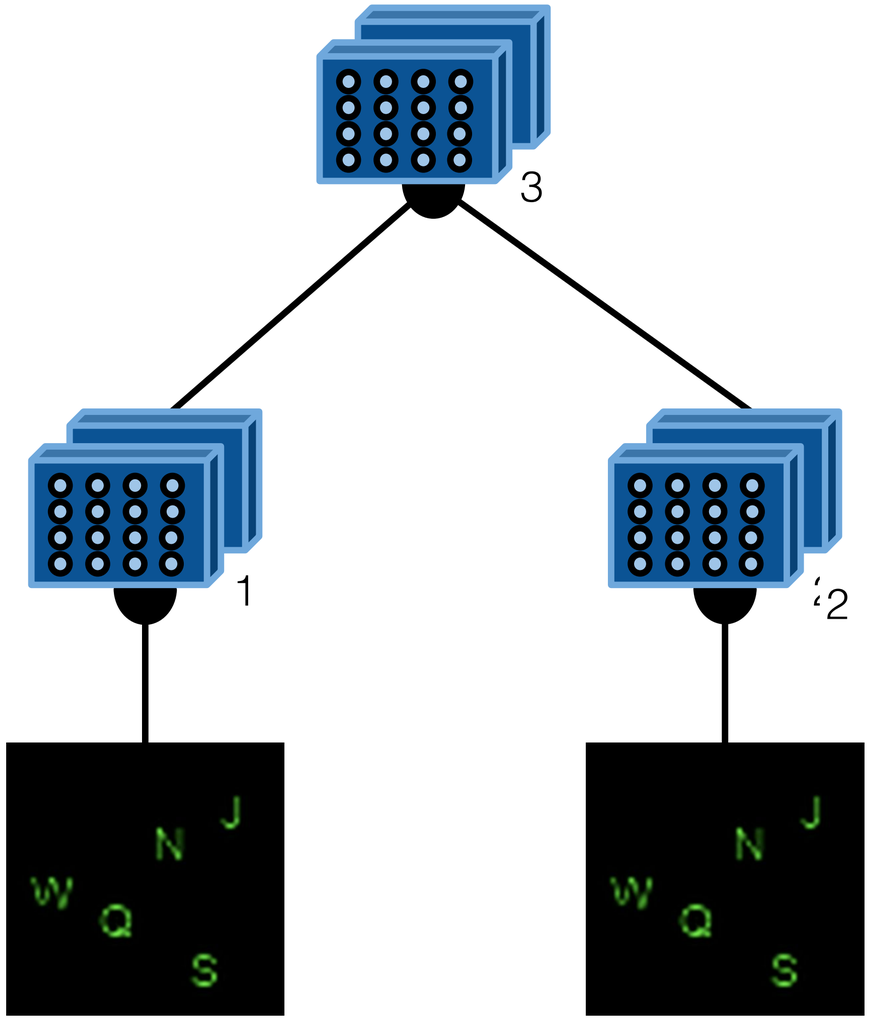
- NMN-Tree generalizes better in the most difficult version of the problem, i.e. when #rhs/lhs = 1.



Experiment II: Importance of NMN tree layout

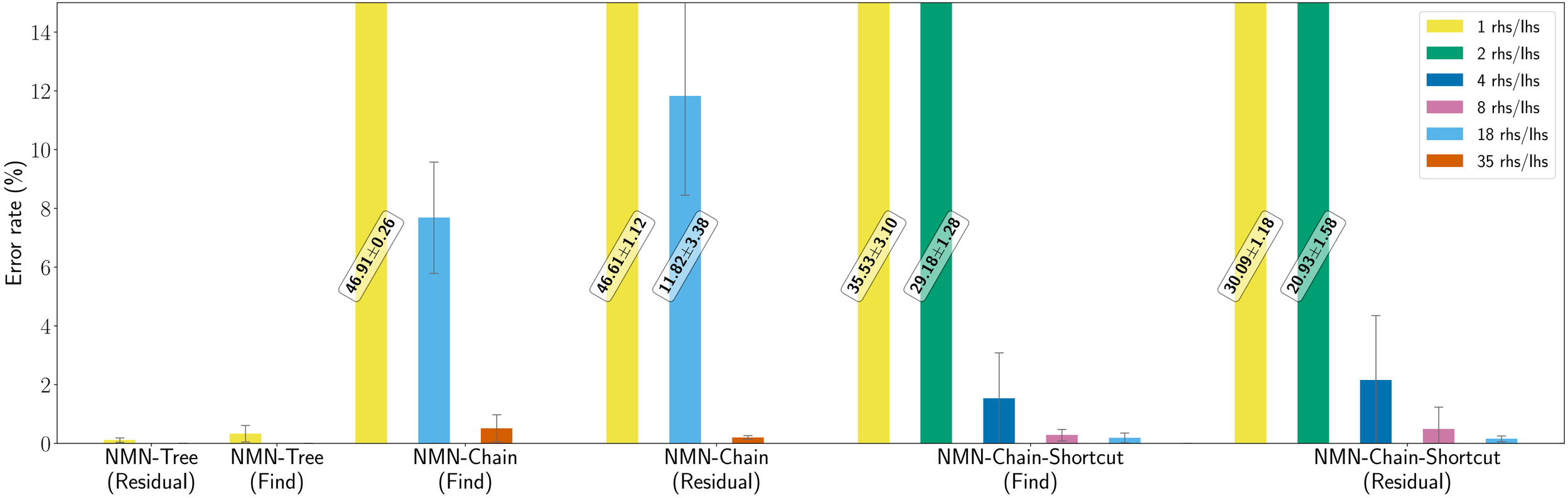


— or —



Experiment II: Importance of tree-structured models

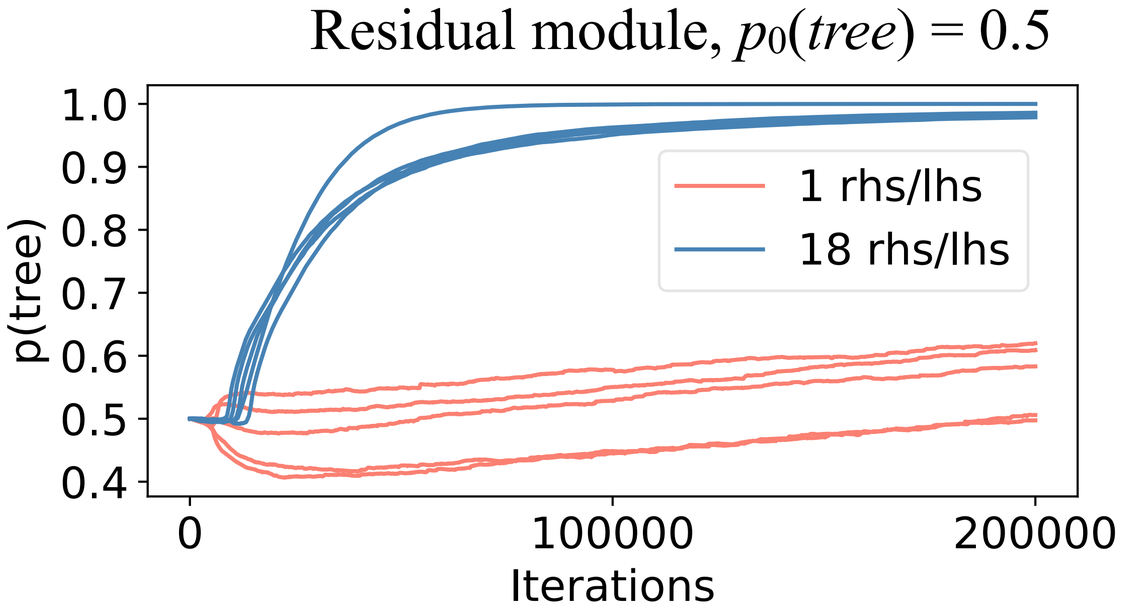
- NMNs only generalize when they are in the “correct” tree structure, irrespective of the module type.



Experiment III: can the NMN layout be induced?

- Evaluate layout induction in an “ideal” setting.
- Consider the layout T as a stochastic latent variable: $T \in \{T_{\text{chain}}, T_{\text{tree}}\}$

$$p(\text{True} \mid x, q) = \sum_{T \in \{T_{\text{chain}}, T_{\text{tree}}\}} p(\text{True} \mid T, x, q) p(T)$$



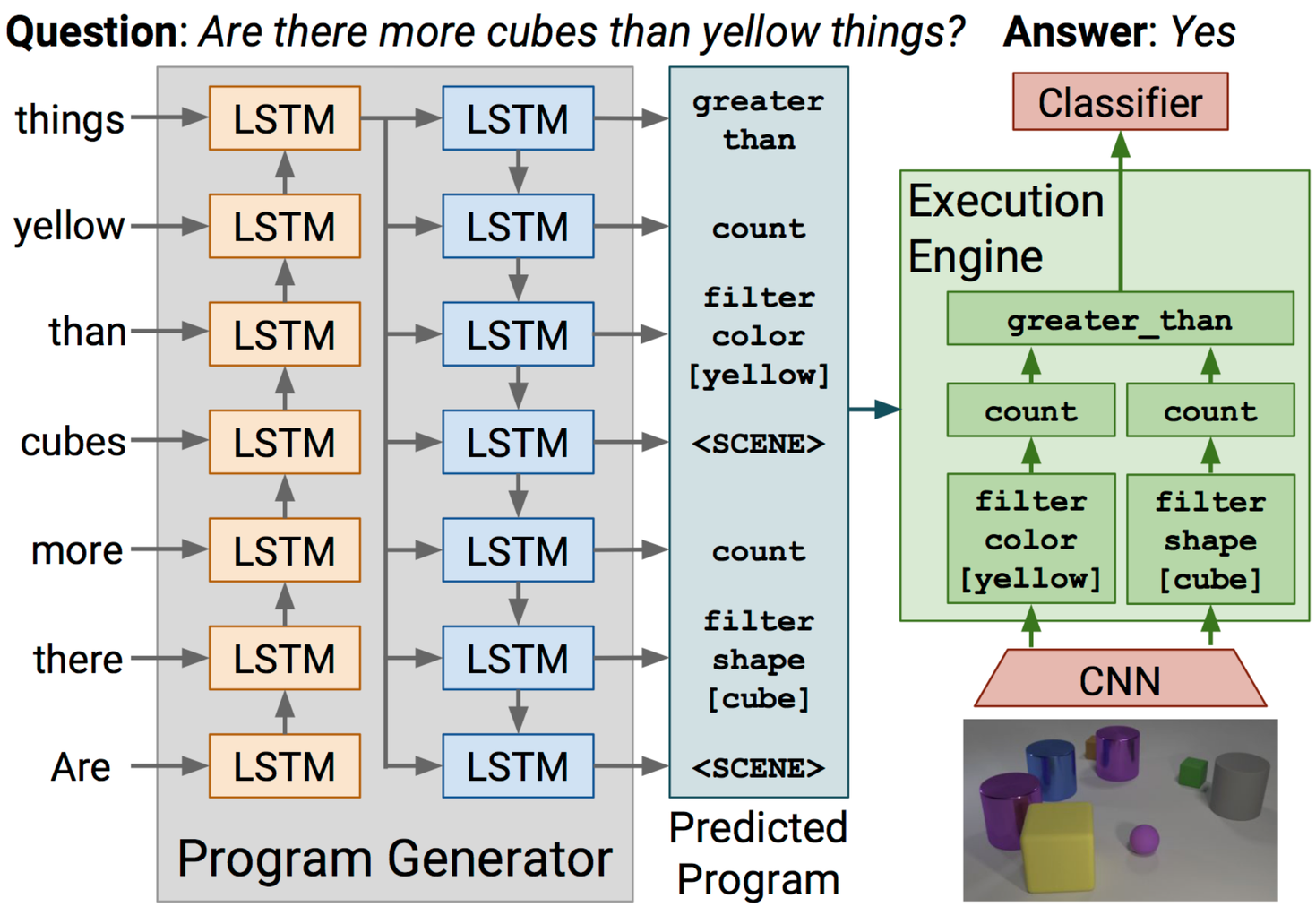
(a) Residual modules

#rhs/lhs	$p_0(tree)$	test acc. (%)	$p_{50K}(tree)$
1	0.1	52.7 ± 2.2	0.003
	0.5	57.0 ± 4.4	0.026
	0.9	99.9 ± 0.1	0.997
18	0.1	100.0 ± 0.0	0.999
	0.5	97.7 ± 5.1	0.999
	0.9	99.1 ± 2.3	0.999

(b) Find modules

#rhs/lhs	$p_0(tree)$	test acc. (%)	$p_{50K}(tree)$
1	0.1	51.2 ± 2.9	0.00
	0.5	93.2 ± 7.1	0.999
	0.9	95.9 ± 1.6	0.999
18	0.1	78.6 ± 20.7	0.2
	0.5	91.6 ± 6.5	0.999
	0.9	97.3 ± 3.4	0.999

NMNs for Systematic Generalization

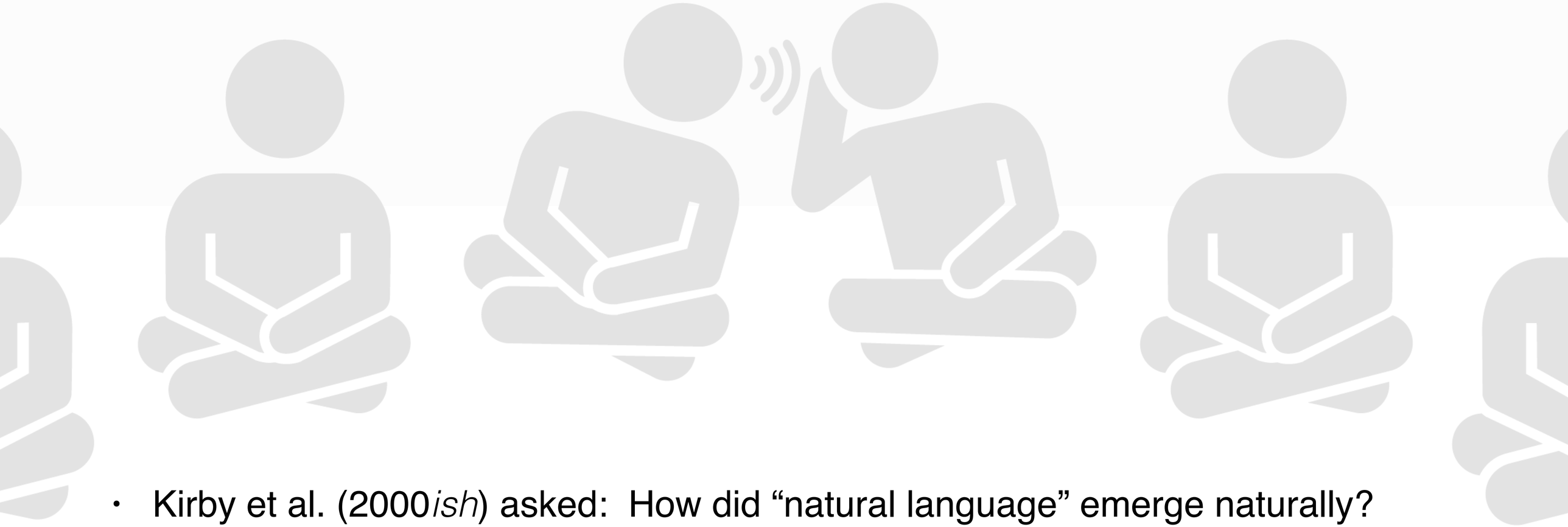


- NMN are inherently modular and compositional.
- NMNs are well-suited for systematic generalization
- Though inferring the correct layout / program is difficult.
  - Training error is not helpful.

\- Bahdanau et al. (2019)

Johnson et al.
Inferring and Executing Programs for Visual Reasoning.
ICCV 2017.

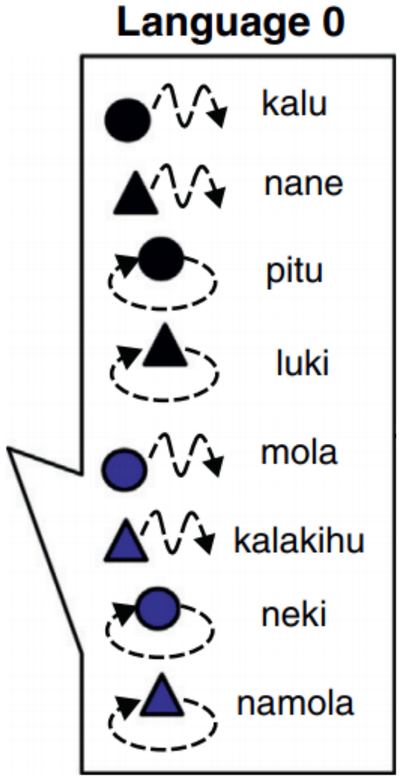
Emergence of Natural Language



- Kirby et al. (2000*ish*) asked: How did “natural language” emerge naturally?
- Where does the regular compositional structure of language come from?

Iterated Learning: human language emergence

Iterated Learning: the compositionally structure of human language emerged through the successive re-learning of language from generation to generation.

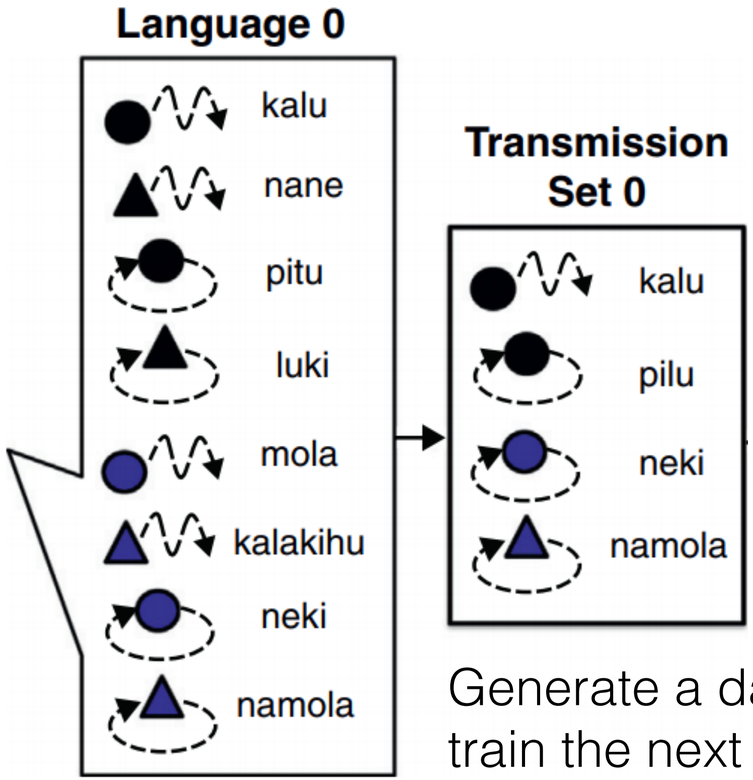


Start with a set of shape/actions

Simon Kirby, Tom Griffiths, and Kenny Smith. “Iterated learning and the evolution of language.” In: *Current opinion in neurobiology* 28 (2014),

Iterated Learning: human language emergence

Iterated Learning: the compositionally structure of human language emerged through the successive re-learning of language from generation to generation.

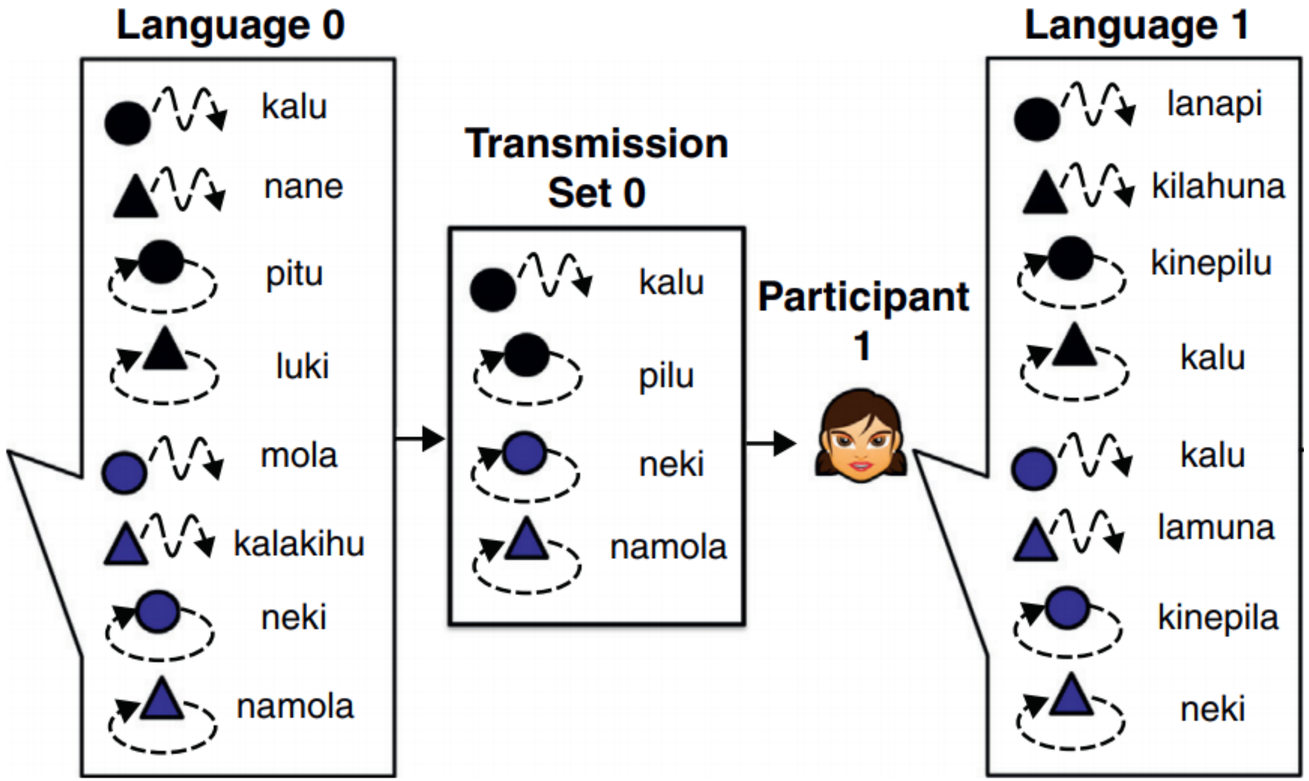


Generate a dataset to train the next generation

Simon Kirby, Tom Griffiths, and Kenny Smith. “Iterated learning and the evolution of language.” In: *Current opinion in neurobiology* 28 (2014),

Iterated Learning: human language emergence

Iterated Learning: the compositionally structure of human language emerged through the successive re-learning of language from generation to generation.

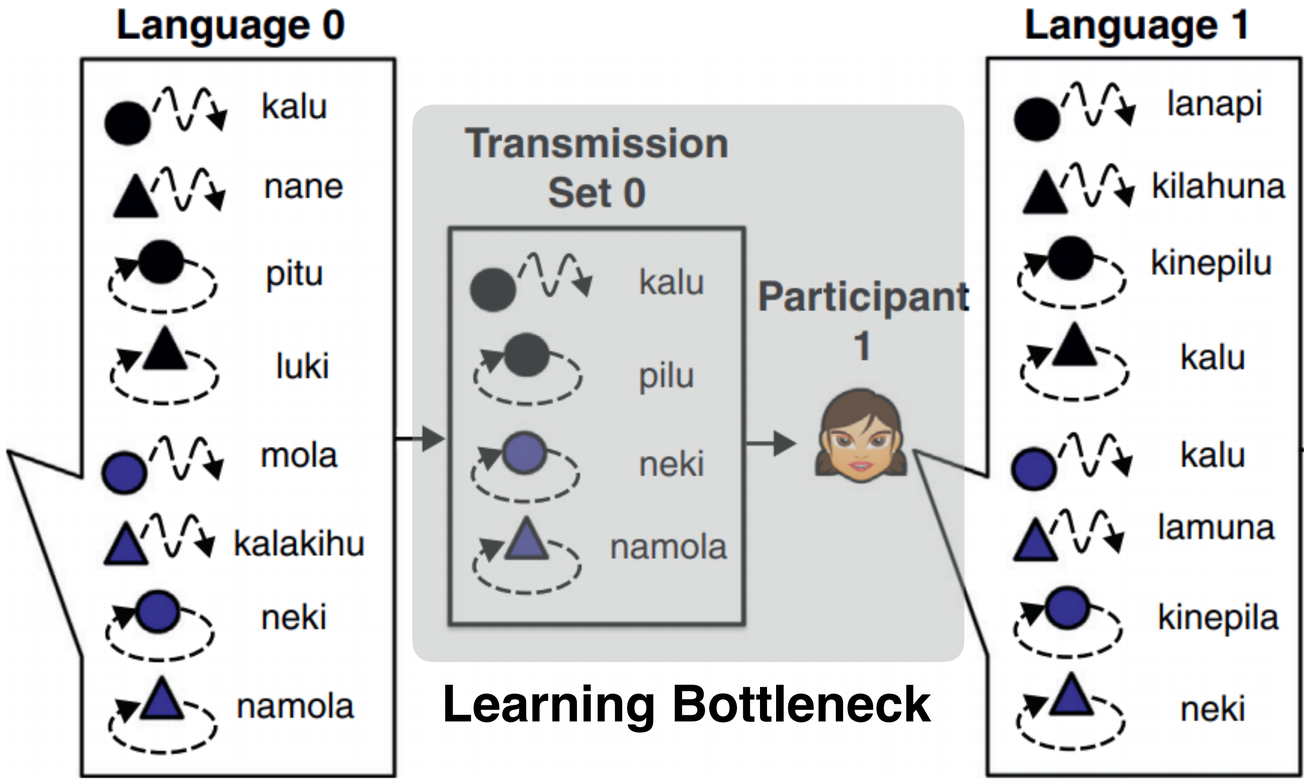


Training the next generation from the limited data

Simon Kirby, Tom Griffiths, and Kenny Smith. “Iterated learning and the evolution of language.” In: *Current opinion in neurobiology* 28 (2014),

Iterated Learning: human language emergence

Iterated Learning: the compositionally structure of human language emerged through the successive re-learning of language from generation to generation.

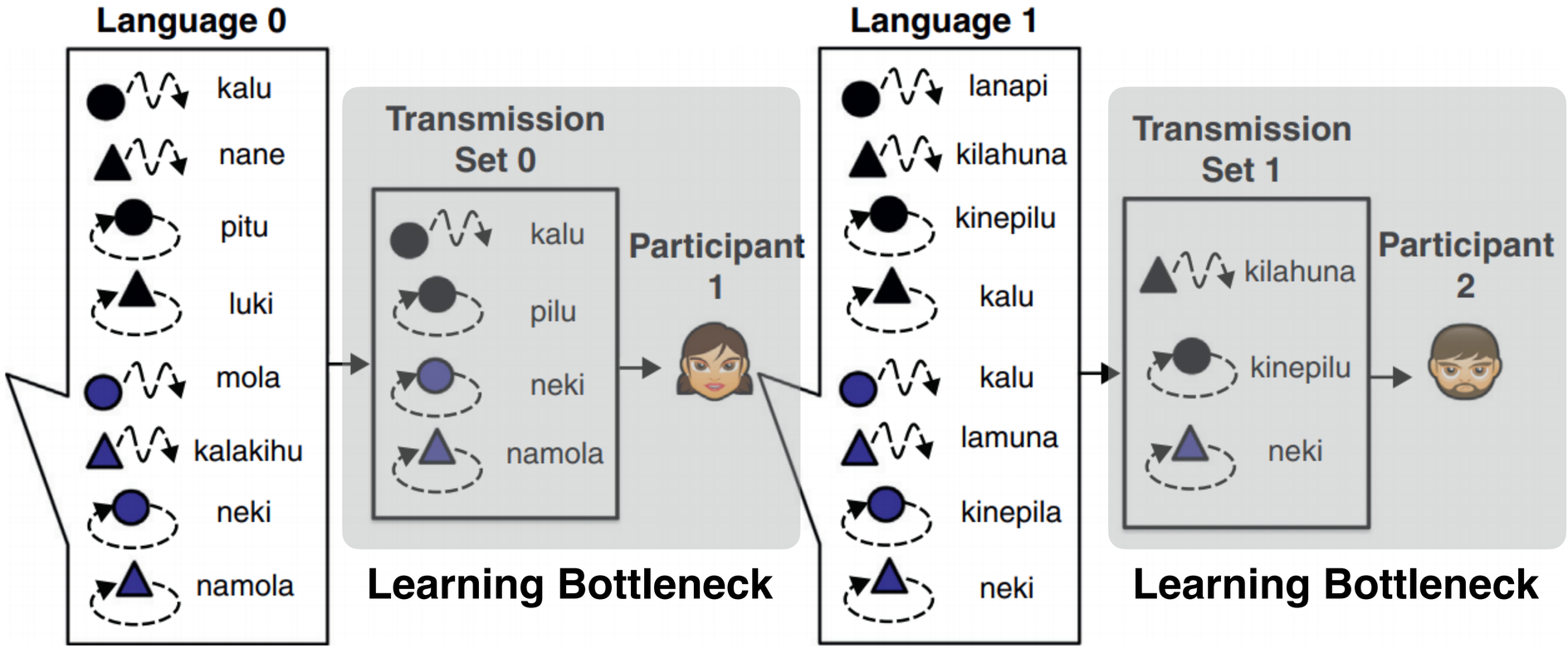


Training the next generation from the limited data

Simon Kirby, Tom Griffiths, and Kenny Smith. "Iterated learning and the evolution of language." In: *Current opinion in neurobiology* 28 (2014),

Iterated Learning: human language emergence

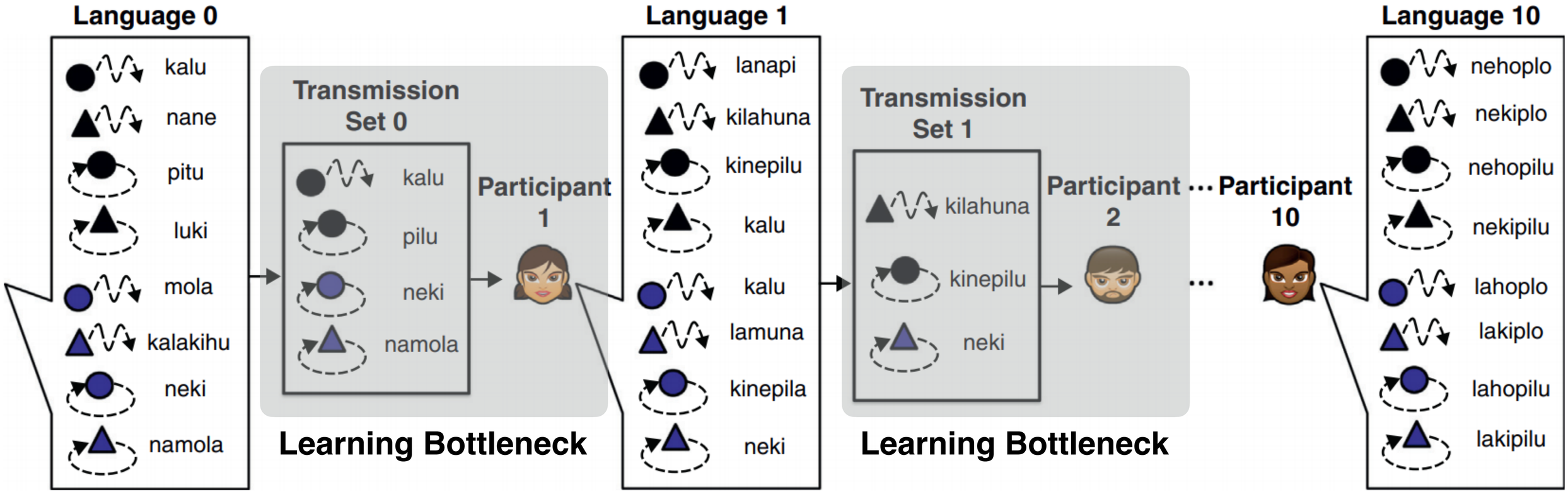
Iterated Learning: the compositionally structure of human language emerged through the successive re-learning of language from generation to generation.



Simon Kirby, Tom Griffiths, and Kenny Smith. "Iterated learning and the evolution of language." In: *Current opinion in neurobiology* 28 (2014),

Iterated Learning: human language emergence

Iterated Learning: the compositionally structure of human language emerged through the successive re-learning of language from generation to generation.



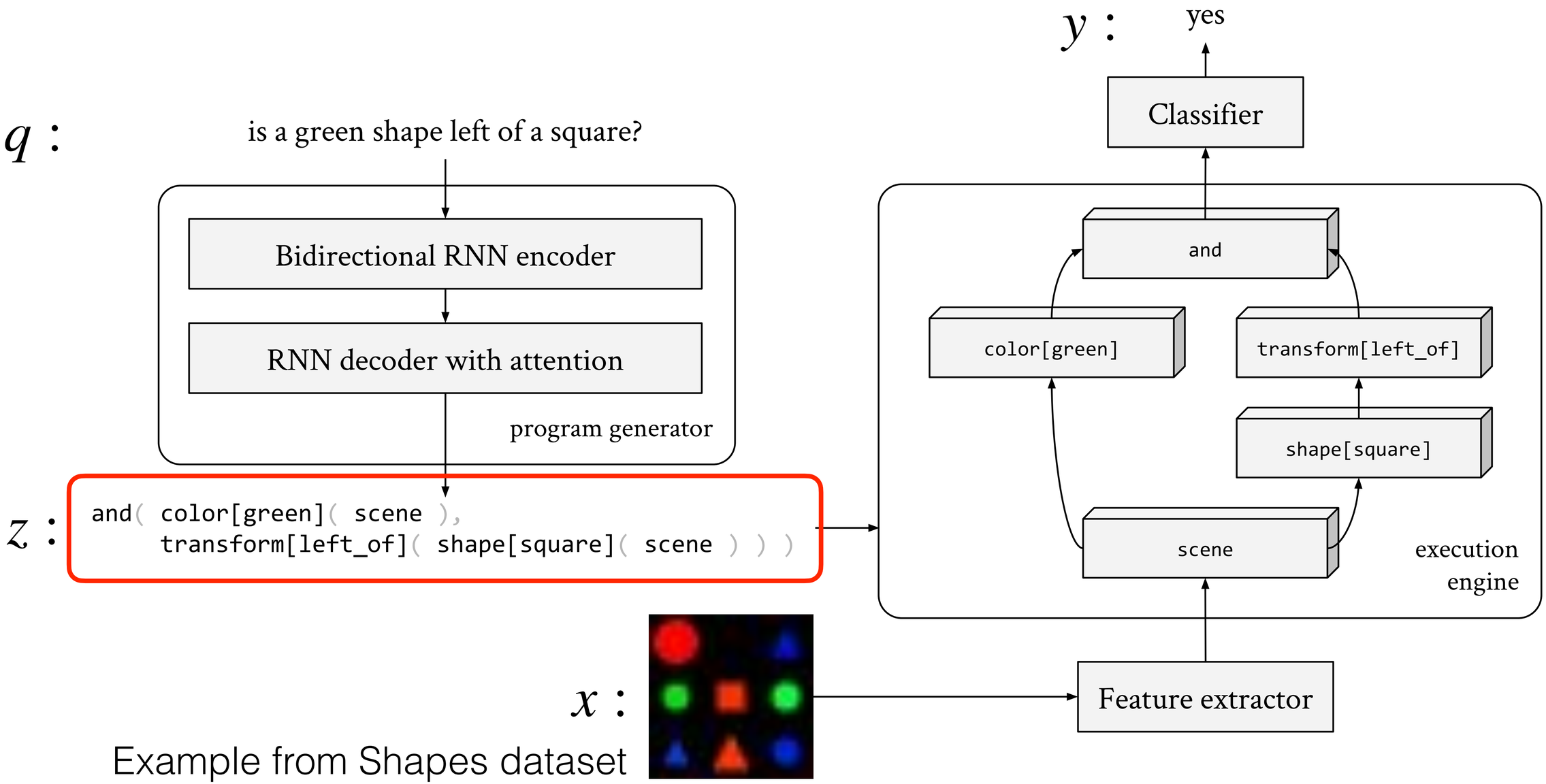
Simon Kirby, Tom Griffiths, and Kenny Smith. "Iterated learning and the evolution of language." In: *Current opinion in neurobiology* 28 (2014),

SHAPES (Andreas et al, 2016) & SHAPES-SyGeT (Vani et al, 2020)

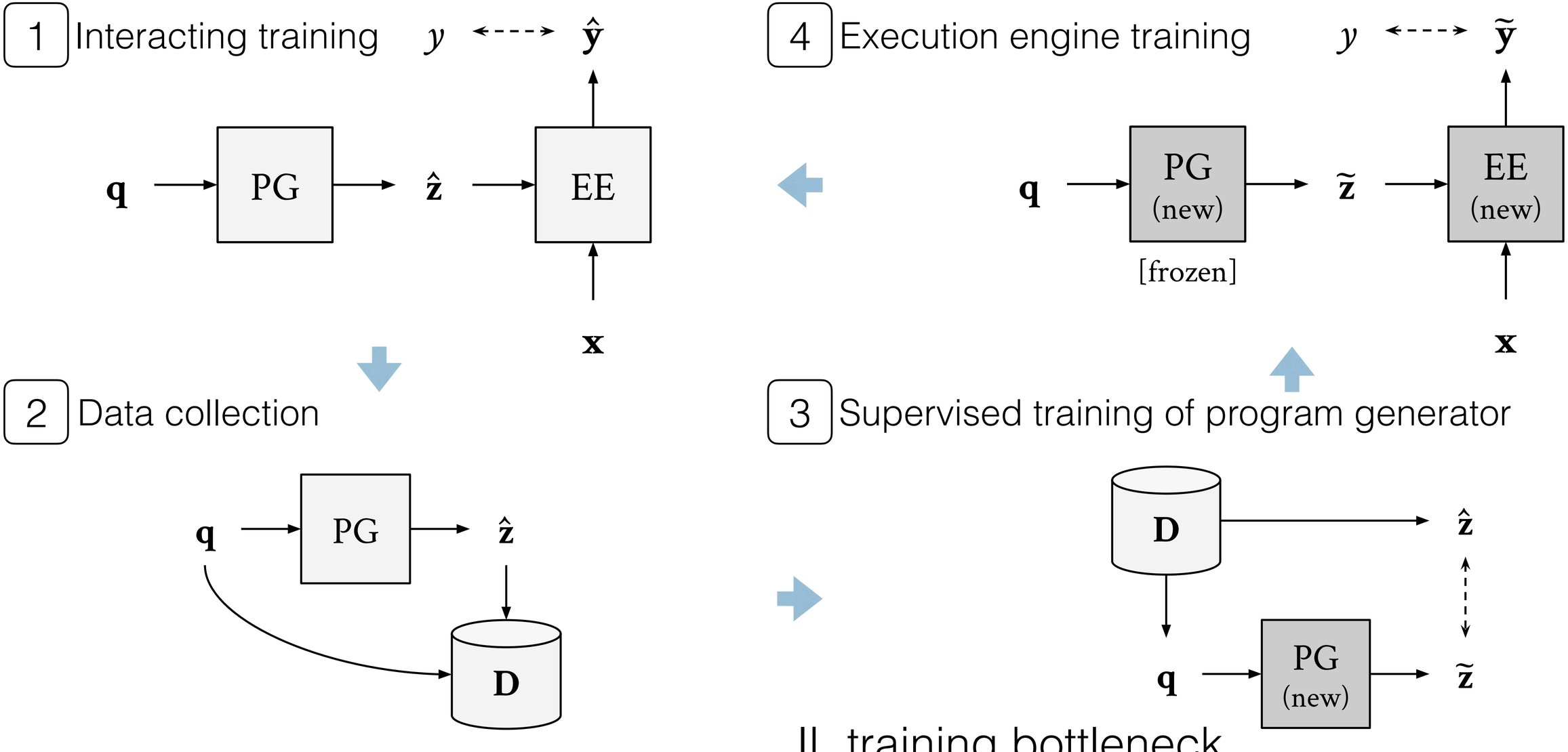
- SHAPES is a visually simple Visual Question-Answer (VQA) dataset
- Shapes-SyGeT dataset:
  - 7 training set question templates and 5 systematic-generalization set question templates.

Q1 (SHAPES-SyGeT train): Is a red shape above a green shape?
Q2 (SHAPES-SyGeT train): Is a *circle* left of below a *square*?
Q3 (SHAPES-SyGeT eval): Is a red shape left of below a green shape?

Neural Module Networks



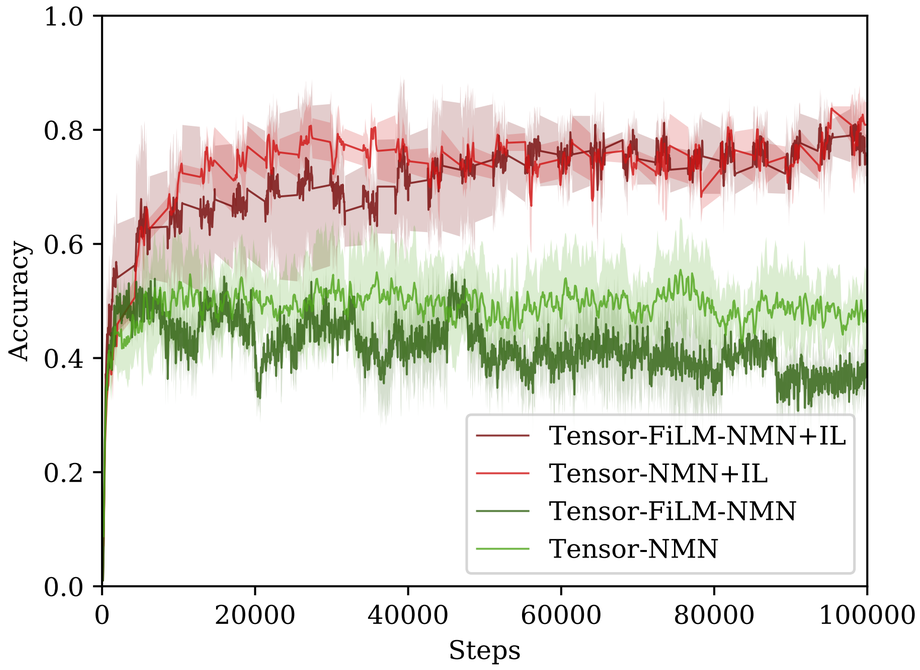
Phases of Iterated Learning of NMN



IL training bottleneck

Iterated learning improves IID / systematic generalization

Shapes dataset: IID Program Accuracy



Test answer / program accuracy improved by Iterated Learning.

Shapes-SyGeT dataset: 7 training templates and 5 systematic-generalization templates.

Model	Val-IID		Val-OOD	
FiLM (Perez et al., 2018)	0.720 ± 0.01		$\mathbf{0.609 \pm 0.01}$	
MAC (Hudson & Manning, 2018)	$\mathbf{0.730 \pm 0.01}$		0.605 ± 0.01	
	#GT 20	**#GT 135**	**#GT 20**	**#GT 135**
`Tensor-NMN`	0.645 ± 0.01	0.700 ± 0.01	0.616 ± 0.01	0.641 ± 0.003
`Tensor-NMN+IL`	0.756 ± 0.07	0.763 ± 0.04	0.648 ± 0.02	0.661 ± 0.02
`Tensor-FiLM-NMN`	0.649 ± 0.02	0.851 ± 0.01	0.605 ± 0.01	0.692 ± 0.06
`Tensor-FiLM-NMN+IL`	$\mathbf{0.954 \pm 0.07}$	$\mathbf{1.000 \pm 0.00}$	$\mathbf{0.858 \pm 0.15}$	$\mathbf{0.971 \pm 0.02}$

Systematic generalization is improved by Iterated Learning.

Ankit Vani

Max Schwarzer

Eeshan Dhekane

Yuchen Lu

Open Questions 1/2

- Iterated learning is an interesting training strategy, but ….
  - ➡ Why does it work?
  - ➡ Can we adapt it to a wider range of tasks?
  - ➡ Can we scale it up?
- What should be our basis of comparison of different methods?
  - ➡ How do we tradeoff IID generalization versus OOD generalization?

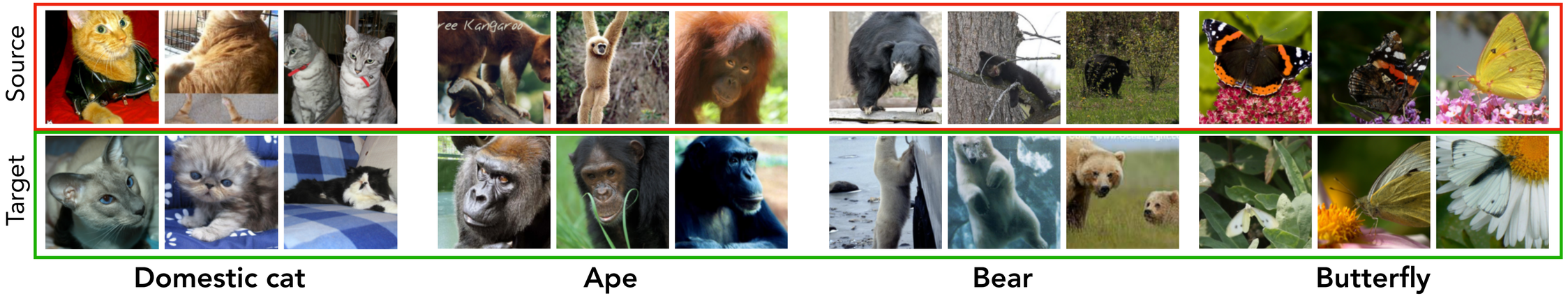
Open Questions 2/2

How much of OOD generalization can be considered systematic generalization?

- Generalizing from some dog breeds to other dog breeds? … Probably yes.

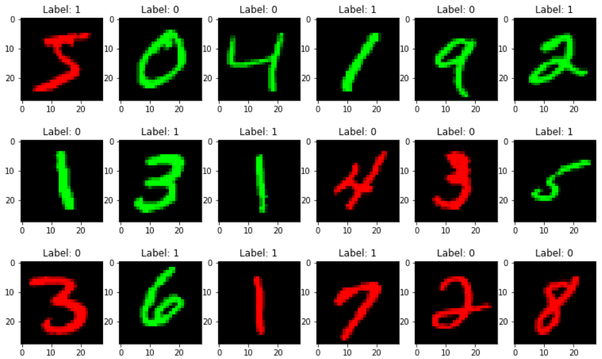
LIVING-17

BREEDS benchmark (Santurkar et al 2020)



- IRM-style coloured-MNIST (w/ label-correlated colours)

Training set:



Test set: