

WELCOME TO THE

# UCL x DeepMind lecture series

In this lecture series, leading research scientists from leading AI research lab, DeepMind, will give 12 lectures on an exciting selection of topics in Deep Learning, ranging from the fundamentals of training neural networks via advanced ideas around memory, attention, and generative modelling to the important topic of responsible innovation.

Please join us for a deep dive lecture series into Deep Learning!

**#UCLxDeepMind**



TODAY'S SPEAKERS

# Irina Higgins + Mihaela Rosca



Irina is a Research Scientist at DeepMind, where her work aims to use multi-disciplinary insights from fields like neuroscience and physics to advance general artificial intelligence through improved representation learning.

Mihaela Rosca is a Research Engineer at DeepMind and a PhD student at UCL, focusing on generative models research and probabilistic modelling, from variational inference to generative adversarial networks and reinforcement learning.





TODAY'S LECTURE

# Frontiers in Deep Learning: Unsupervised Representation Learning

Unsupervised learning is one of the three major branches of machine learning (along with supervised learning and reinforcement learning). It is also arguably the least developed branch. Its goal is to find a parsimonious description of the input data by uncovering and exploiting its hidden structures. This is presumed to be more reminiscent of how the brain learns compared to supervised learning. Furthermore, it is hypothesised that the representations discovered through unsupervised learning may alleviate many known problems with deep supervised and reinforcement learning. However, lacking an explicit ground truth goal to optimise towards, developmental progress in unsupervised learning has been slow. In this talk we will overview the historical role of unsupervised representation learning and difficulties with developing and evaluating such algorithms. We will then take a multidisciplinary approach to think about what might make a good representation and why, before doing a broad overview of the current state of the art approaches to unsupervised representation learning.



DeepMind

# Frontiers in Deep Learning: Unsupervised Representation Learning

Irina Higgins & Mihaela Rosca

UCL x DeepMind Lectures



# Plan for this Lecture

Want to learn more?



**01**

What is unsupervised learning?

**02**

Why is it important?

**03**

What makes a good representation?

**04**

Evaluating the merit of a representation

**05**

Techniques & Applications

**06**

Future

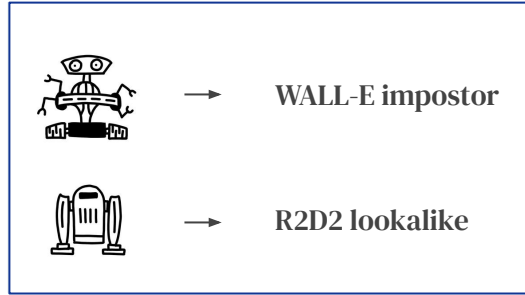


# 1

# What is unsupervised learning?

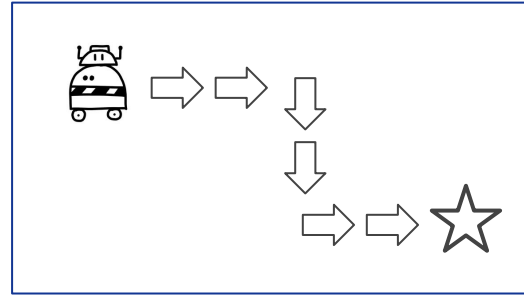


# Bird's eye view



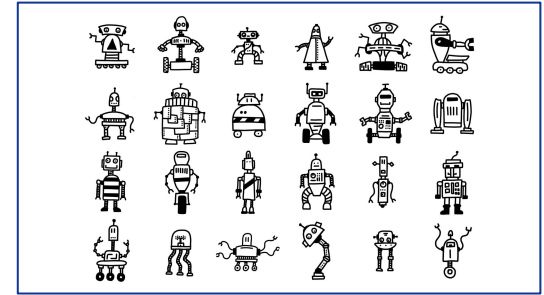
## Supervised Learning

Learn mapping from given input to given output.



## Reinforcement learning

Learn which action will give you more reward in the future.



## Unsupervised learning

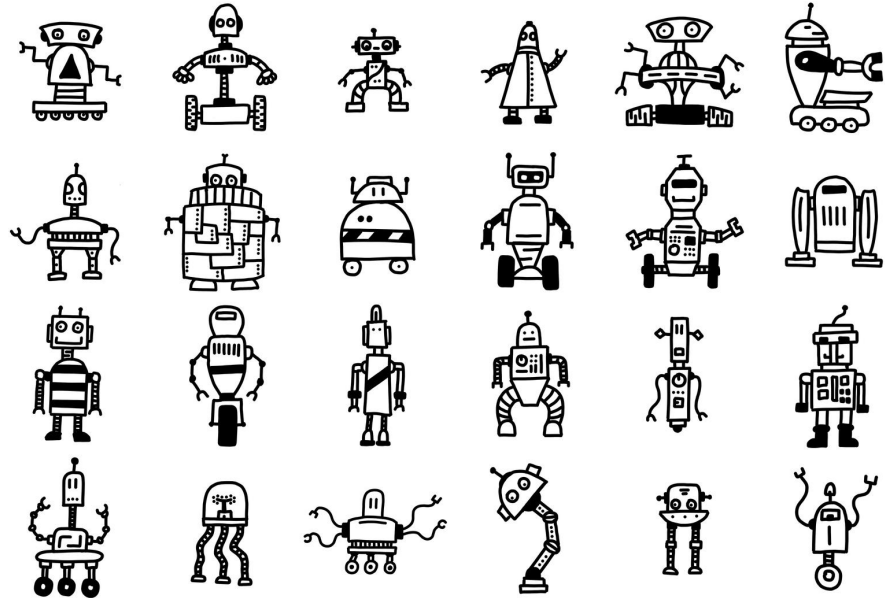
Find structure in provided data.

No teaching signals (labels or rewards).



# Unsupervised learning...

→ Do we need it?



→ How do we evaluate it?

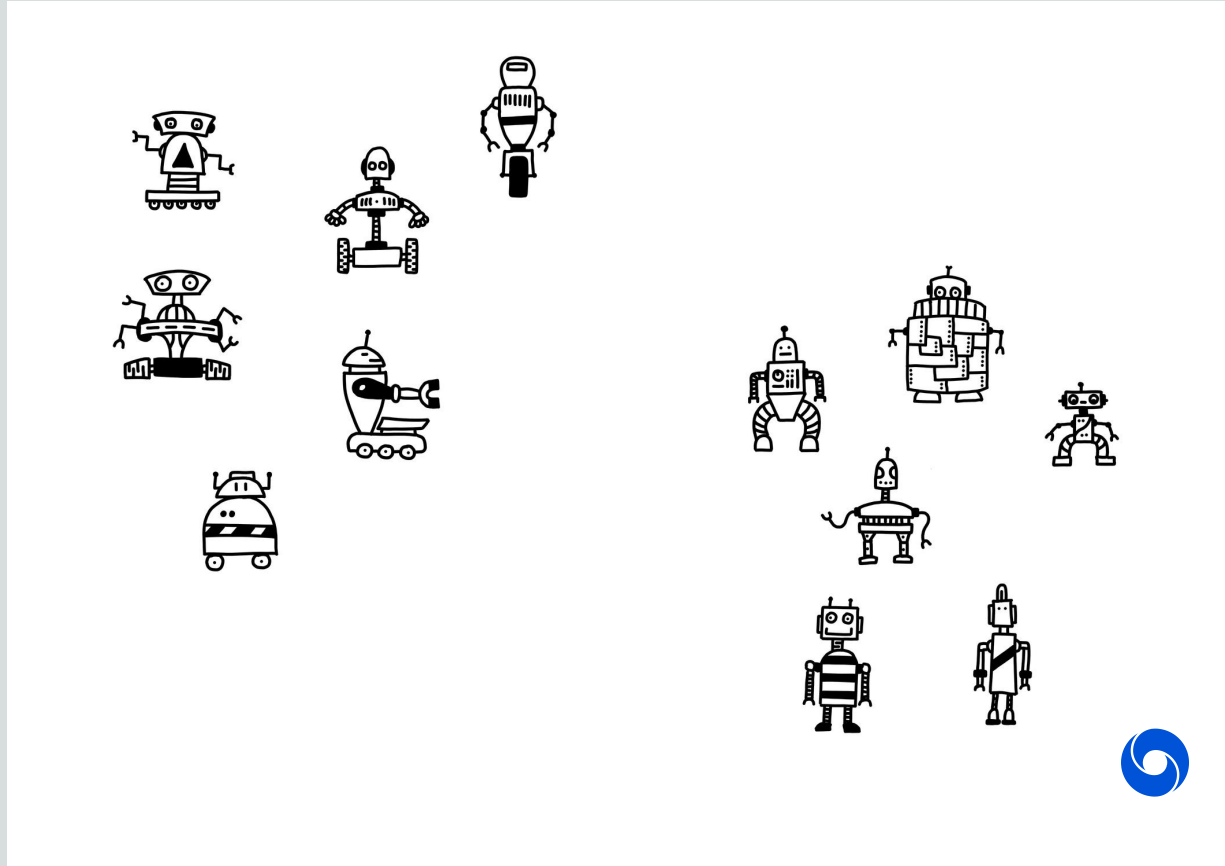




# Unsupervised learning...

→ Do we need it?

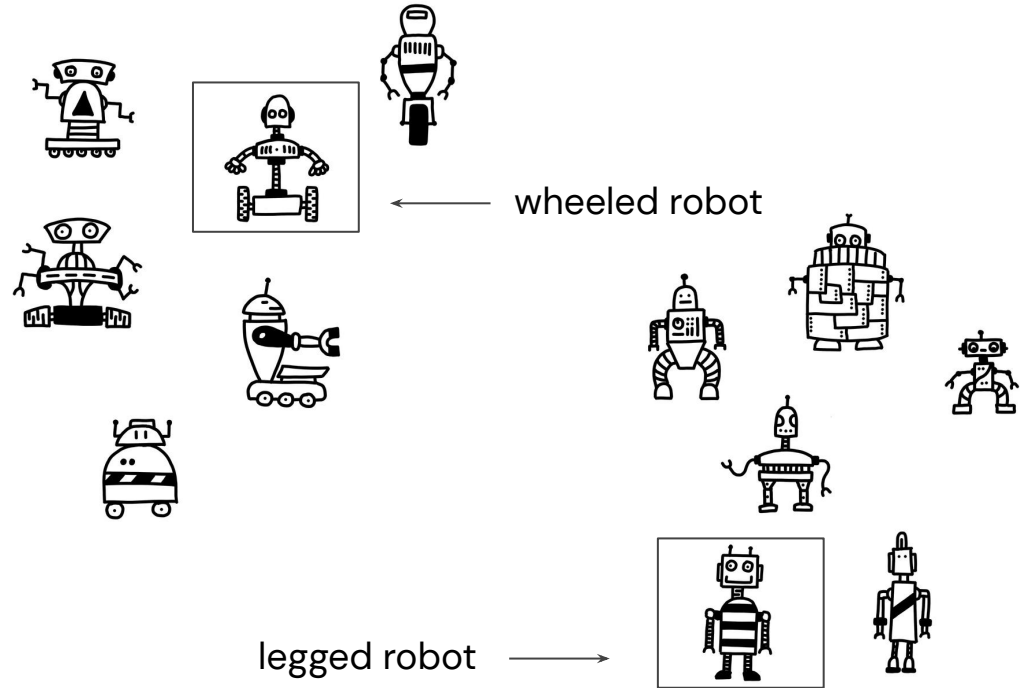
- Clustering



# Unsupervised learning...

→ Do we need it?

- Clustering

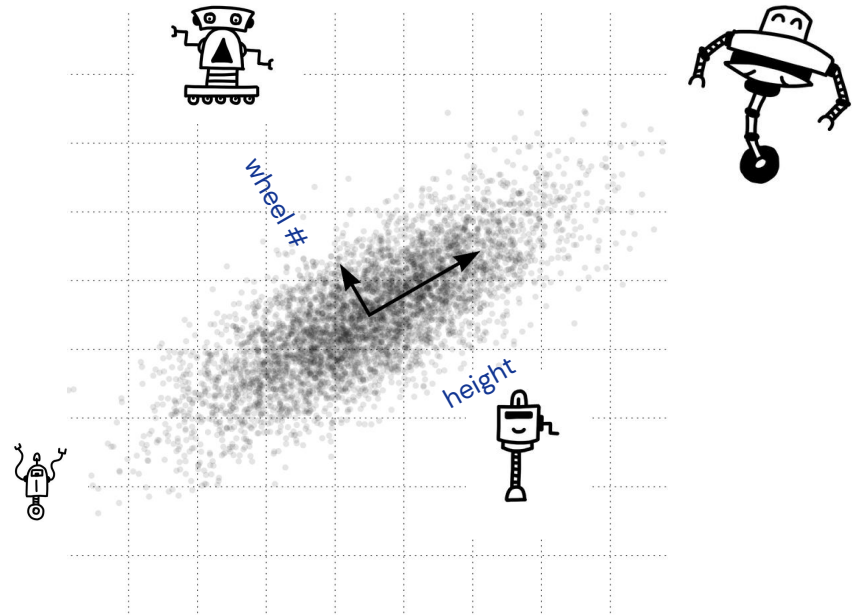


# Unsupervised learning...



Do we need it?

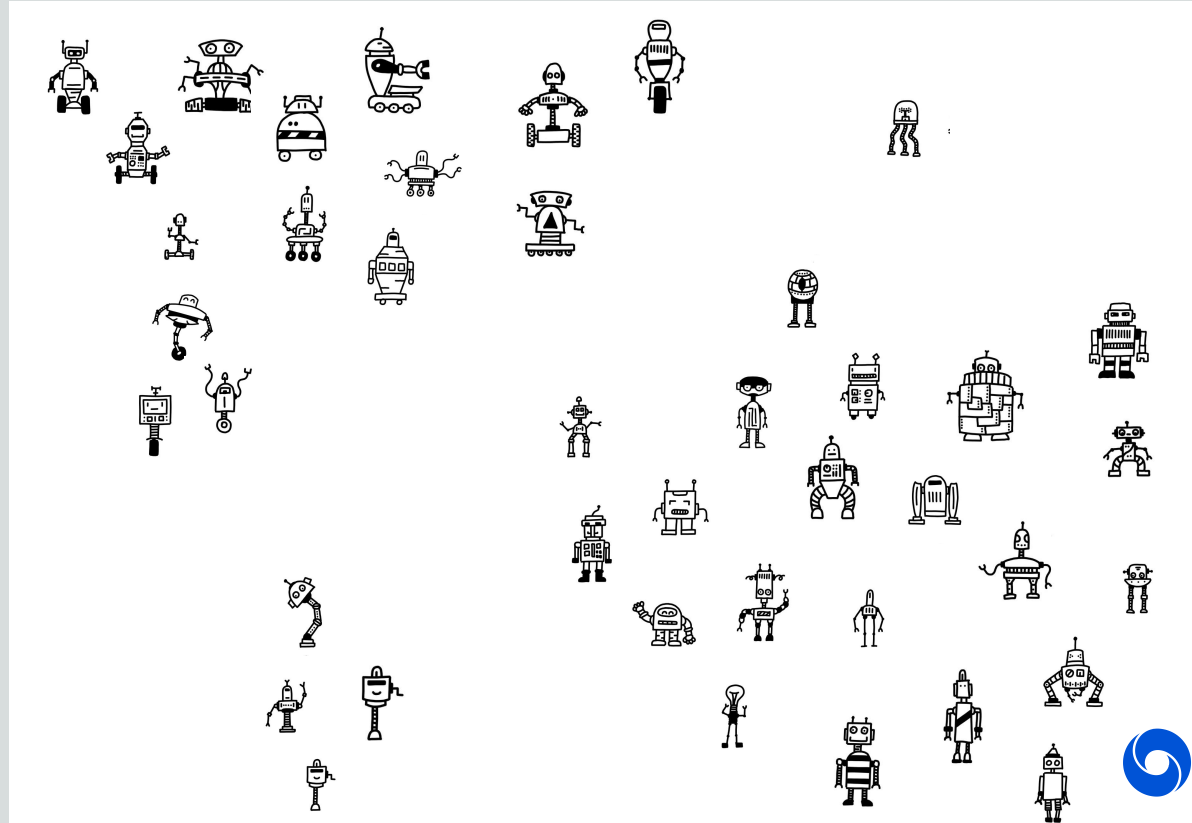
- Clustering
- Dimensionality reduction



# Unsupervised learning...

→ How do we evaluate it?

Finds clusters...  
- by leg type?

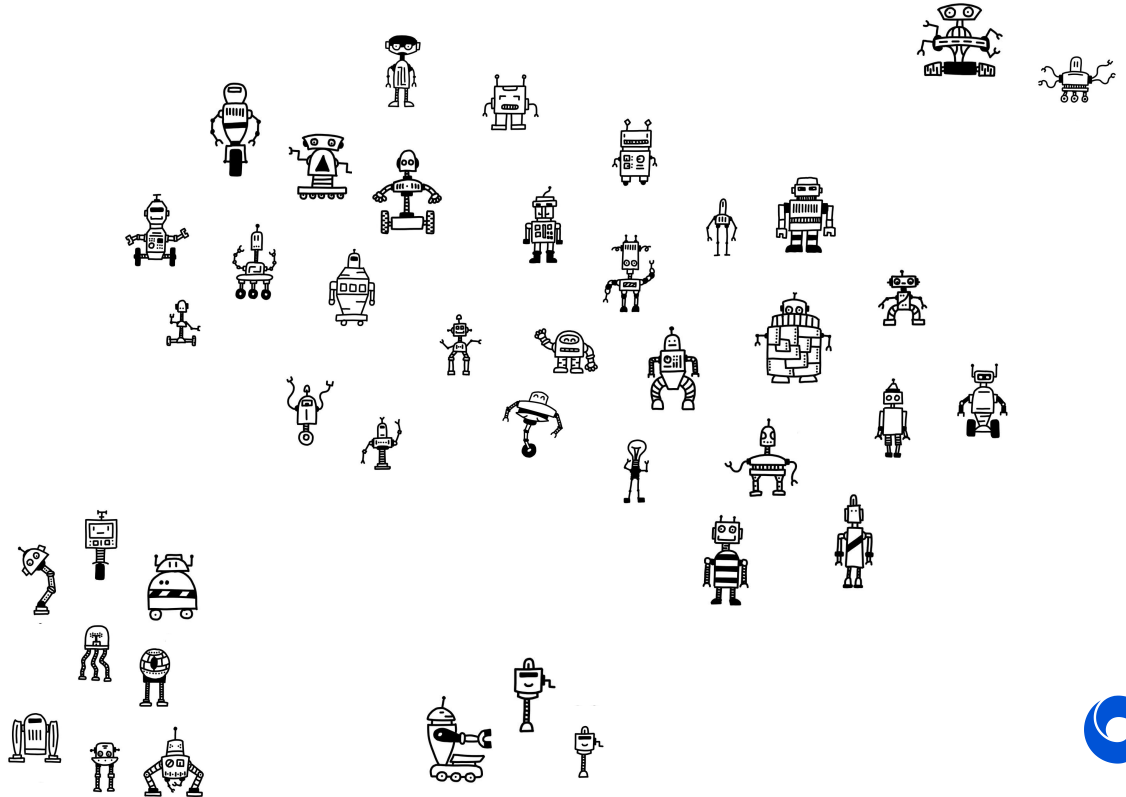


# Unsupervised learning...

→ How do we evaluate it?

Finds clusters...

- by leg type?
- by arm number?

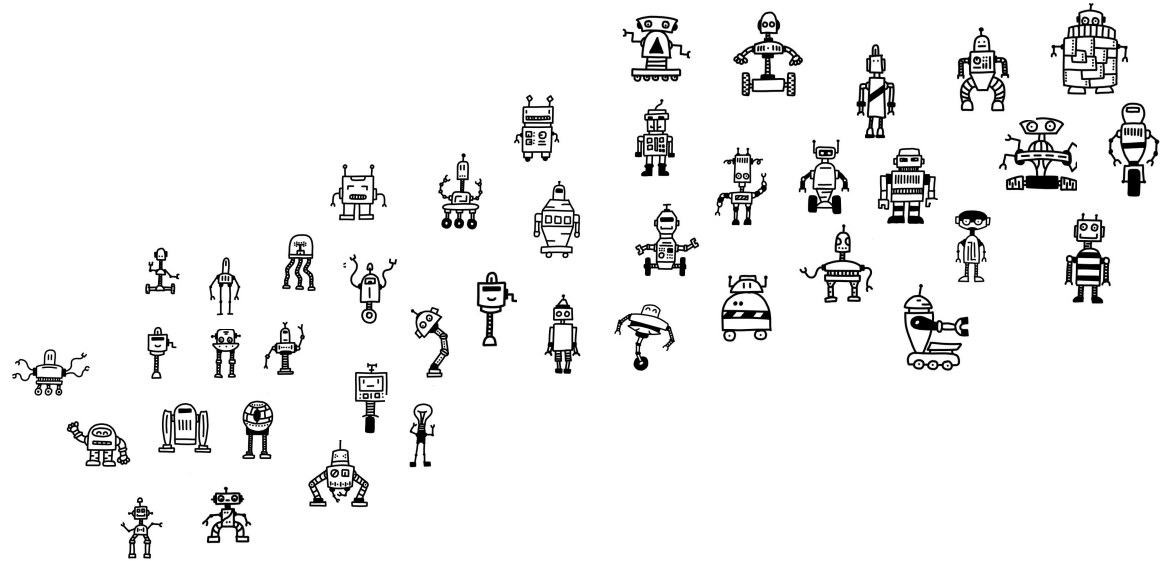


# Unsupervised learning...

→ How do we evaluate it?

Finds clusters...

- by leg type?
- by arm number?
- by height?



# Unsupervised learning...

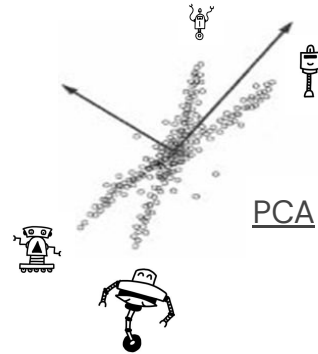
Want to learn more?



Imaging Brain Dynamics Using  
Independent Component  
Analysis, Jung et al, IEEE 2001

→ How do we evaluate it?

Reduce dimensionality...  
- by orthogonality?



# Unsupervised learning...

Want to learn more?

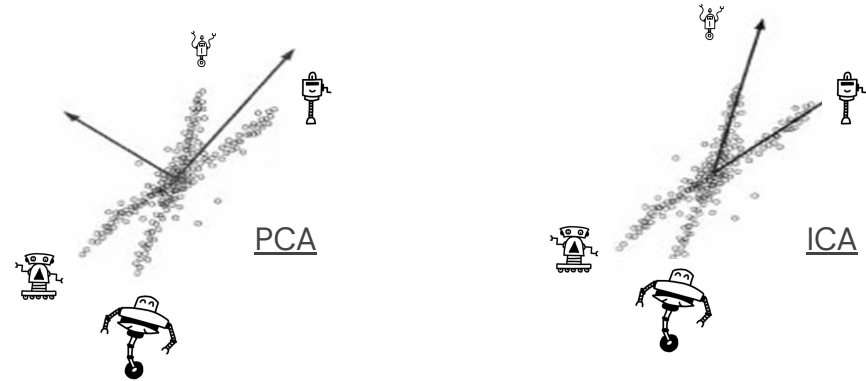


Imaging Brain Dynamics Using  
Independent Component  
Analysis, Jung et al, IEEE 2001

→ How do we evaluate it?

Reduce dimensionality...

- by orthogonality?
- by independence?



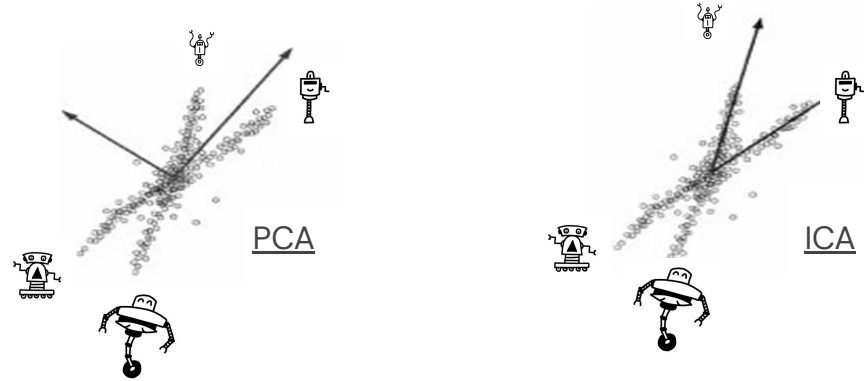


# Unsupervised learning...

→ How do we evaluate it?

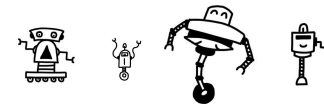
Reduce dimensionality...

- by orthogonality?
- by independence?
- other?



Disentangling

p(height)  
p(leg #)  
p(arm #)  
p(has wheels)



# 2

**Why is  
unsupervised  
learning  
important?**



# History of representation learning

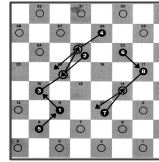
Want to learn more?



Some Studies in Machine Learning  
Using the Game of Checkers,  
Samuel, IBM Journal 1959



Arthur Samuel coins the term “machine learning”



1949



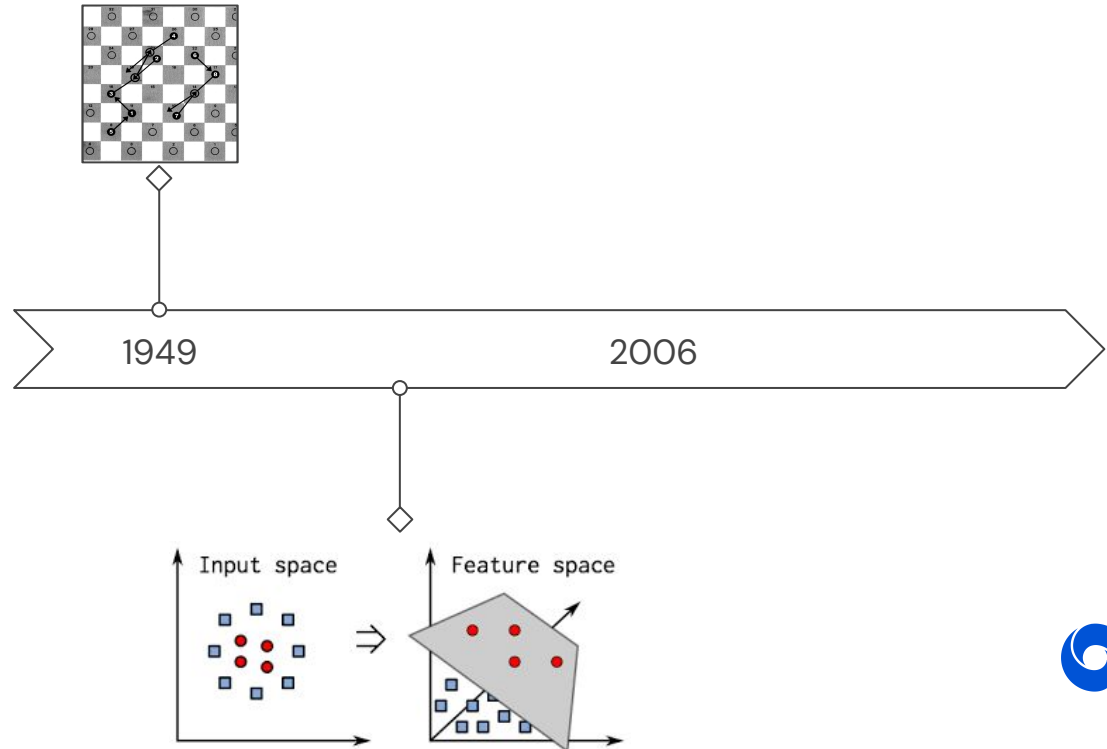
# History of representation learning

Want to learn more?



Kernel Methods in Machine Learning, Hofmann et al, The Annals of Statistics 2008

- Arthur Samuel coins the term “machine learning”
- Feature engineering and kernel methods



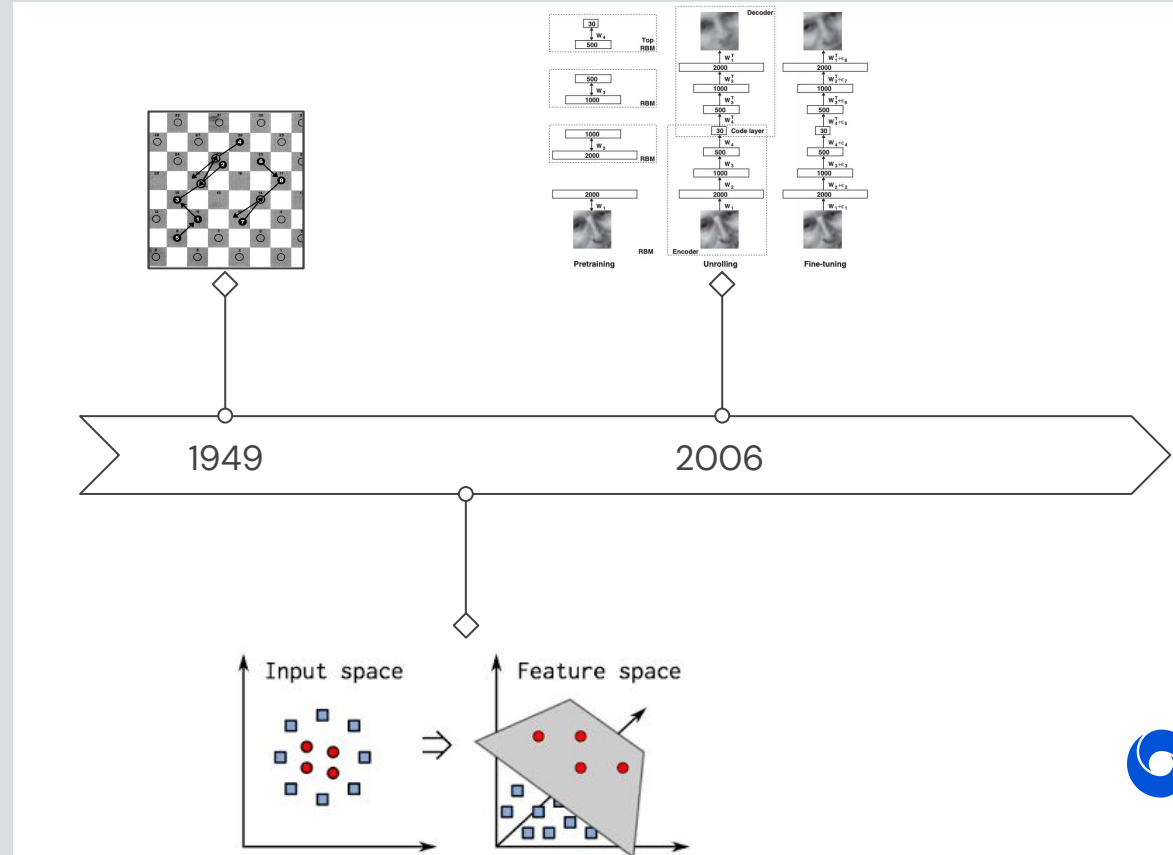
# History of representation learning

Want to learn more?



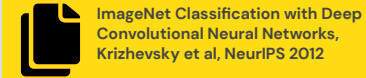
Reducing the Dimensionality of Data with Neural Networks, Hinton and Salakhutdinov, Science 2006

- Arthur Samuel coins the term “machine learning”
- Feature engineering and kernel methods
- Restricted Boltzman Machines used for initialising deep classifiers

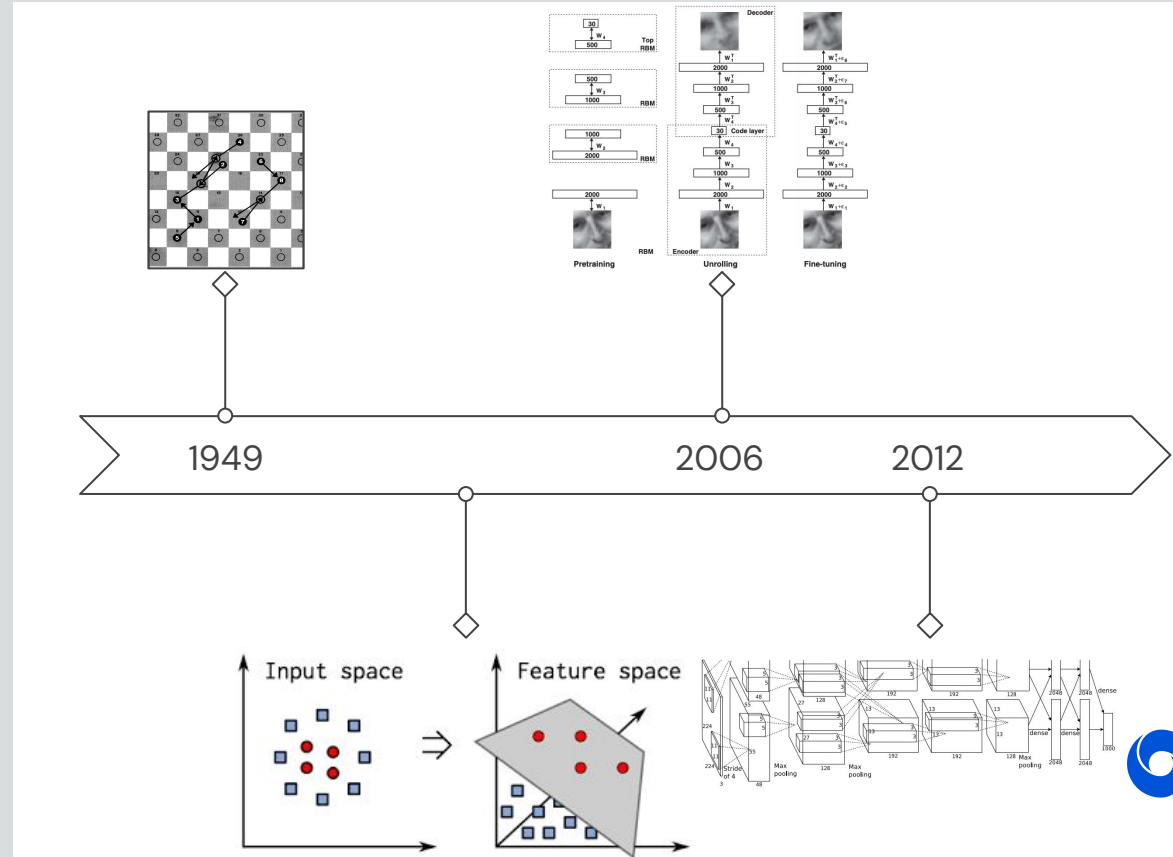


# History of representation learning

Want to learn more?

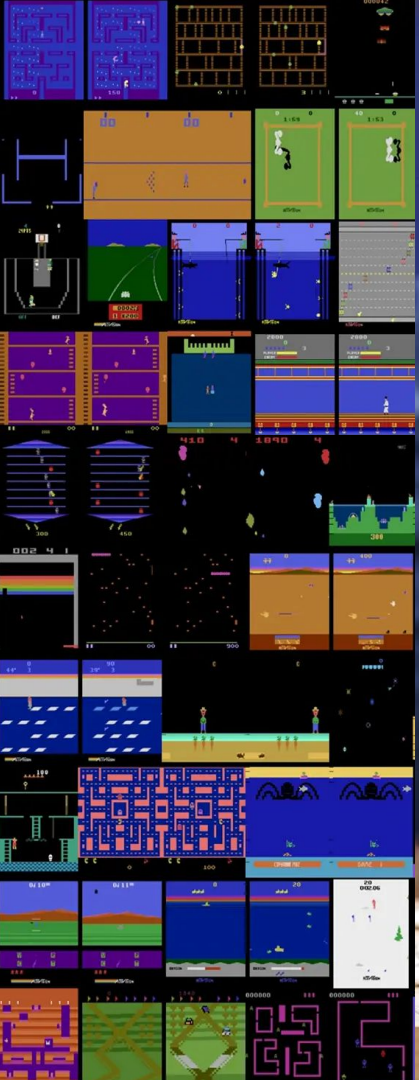
 ImageNet Classification with Deep Convolutional Neural Networks, Krizhevsky et al, NeurIPS 2012

- Arthur Samuel coins the term “machine learning”
- Feature engineering and kernel methods
- Restricted Boltzman Machines used for initialising deep classifiers
- AlexNet wins ImageNet challenge by a large margin with no unsupervised pre-training



**more data**  
+  
**deeper models**  
+  
**better hardware**







# Is machine learning “solved”?

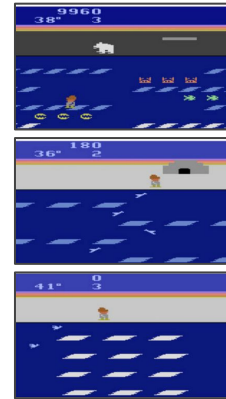
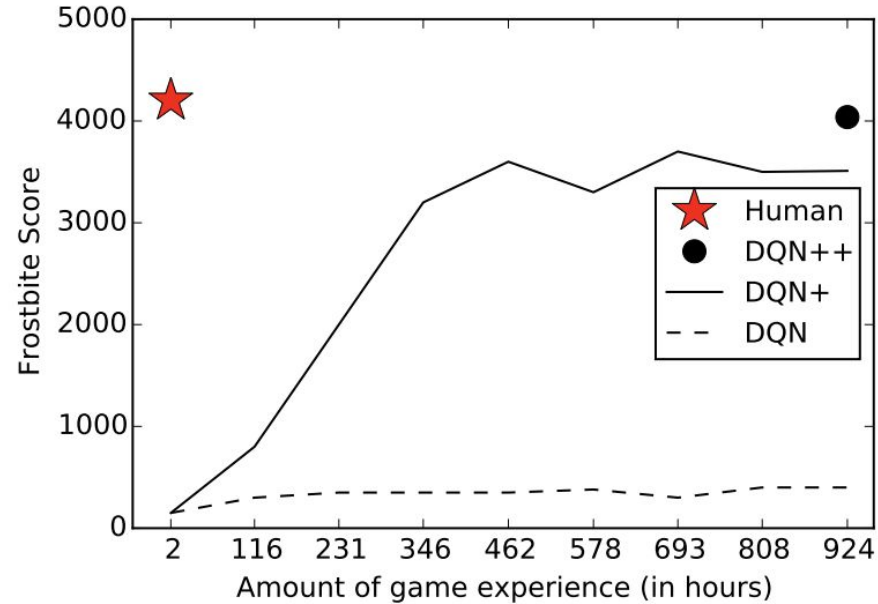
Want to learn more?



Building Machines That Learn and Think Like People, Lake et al, Behavioural and Brain Sciences 2017



Data efficiency



# Is machine learning “solved”?

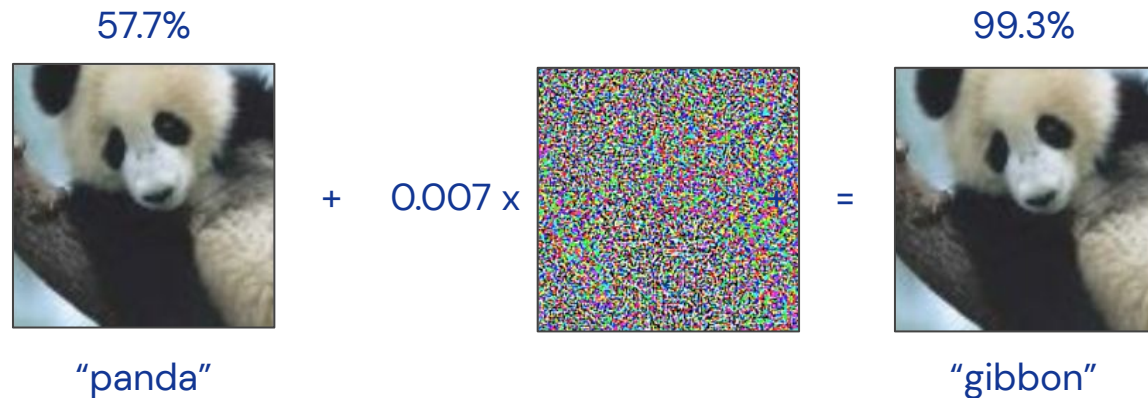
Want to learn more?



Explaining and Harnessing  
Adversarial Examples, Goodfellow  
et al, ICLR 2015

→ Data efficiency

→ Robustness



# Is machine learning “solved”?

Want to learn more?



Why Deep-Learning AIs are so Easy to Fool, Heaven, nature.com 2019

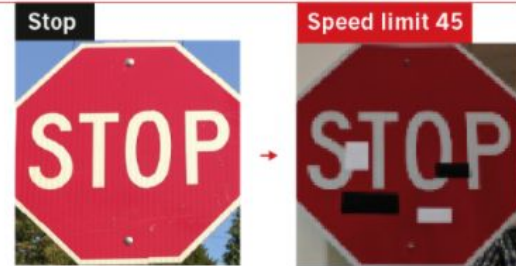
→ Data efficiency

→ Robustness

## FOOLING THE AI

Deep neural networks (DNNs) are brilliant at image recognition — but they can be easily hacked.

These stickers made an artificial-intelligence system read this stop sign as ‘speed limit 45’.



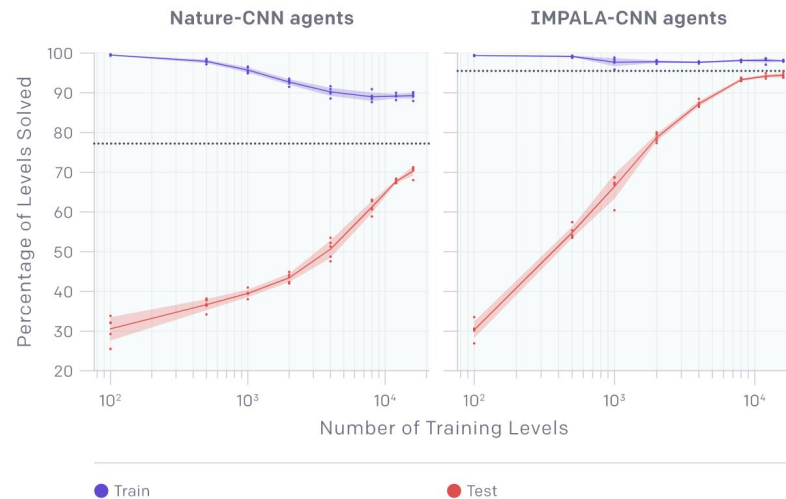
# Is machine learning “solved”?

- Data efficiency
- Robustness
- Generalisation

Want to learn more?



Quantifying Generalization  
in Reinforcement Learning, Cobbe,  
OpenAI Blog 2018



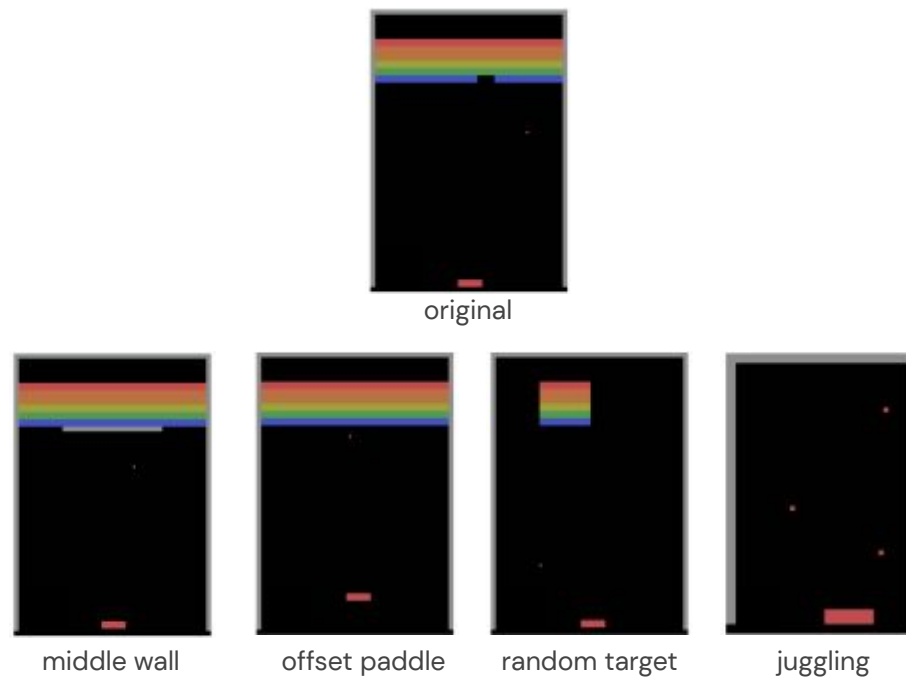
# Is machine learning “solved”?

Want to learn more?



Schema Networks: Zero-shot Transfer with a Generative Causal Model of Intuitive Physics, Kansky et al, ICML 2017

- Data efficiency
- Robustness
- Generalisation
- Transfer



	Standard Breakout	Offset Paddle	Middle Wall	Random Target	Juggling
A3C Image Only	( 36.33 ± 6.17 )	0.60 ± 20.05	9.55 ± 17.44	6.83 ± 5.02	-39.35 ± 14.57



# Is machine learning “solved”?

Want to learn more?



Building Machines That Learn and  
Think Like People, Lake et al,  
Behavioural and Brain Sciences 2017

- Data efficiency
- Robustness
- Generalisation
- Transfer
- “Common sense”



a woman riding a horse on a  
dirt road



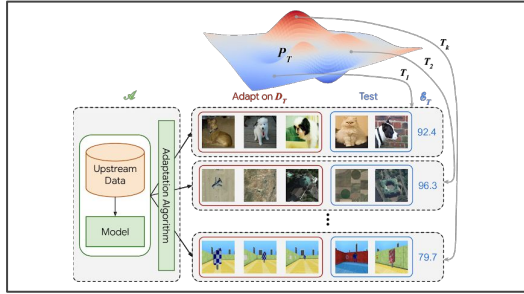
an airplane is parked on the  
tarmac at an airport



a group of people standing on  
top of a beach

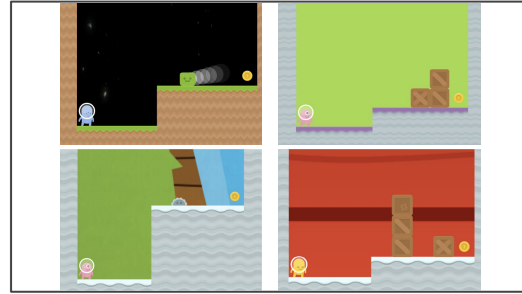


# Solving many tasks efficiently



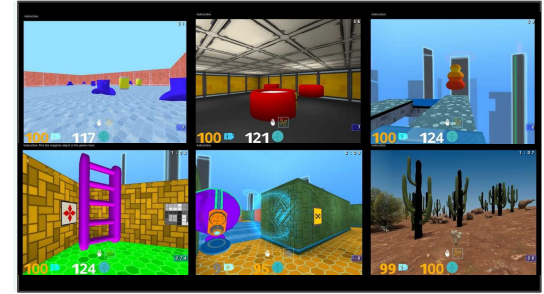
## Visual Task Adaptation Benchmark (Google)

19 visual tasks split into three groups: natural, specialised and structured. Allowance of 1000 adaptation examples per task.



## CoinRun (OpenAI)

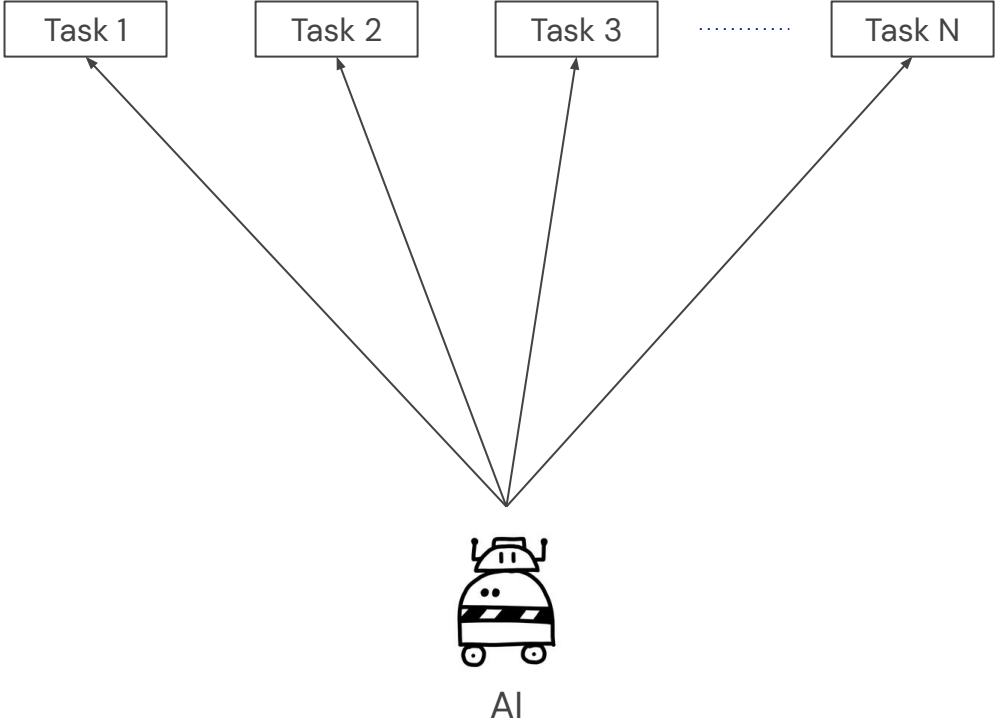
Procedurally generated levels with different degrees of difficulty and a high variability in the game visuals.



## DMLab-30 (DeepMind)

30 varied tasks in a 3D environment, testing navigation, language abilities, multi-agent interactions, long-term planning and more.







# Turing Award winners at AAAI 2020

“

I always knew unsupervised learning was the right thing to do

— Geoff Hinton

“

Basically it's the idea of learning to represent the world before learning a task — and this is what babies do

— Yann LeCun

“

And so if we can build models of the world where we have the right abstractions, where we can pin down those changes to just one or a few variables, then we will be able to adapt to those changes because we don't need as much data, as much observation in order to figure out what has changed.

— Yoshua Bengio



Jérémie Barande / Ecole polytechnique Université Paris-Saclay / CC BY-SA 2.0

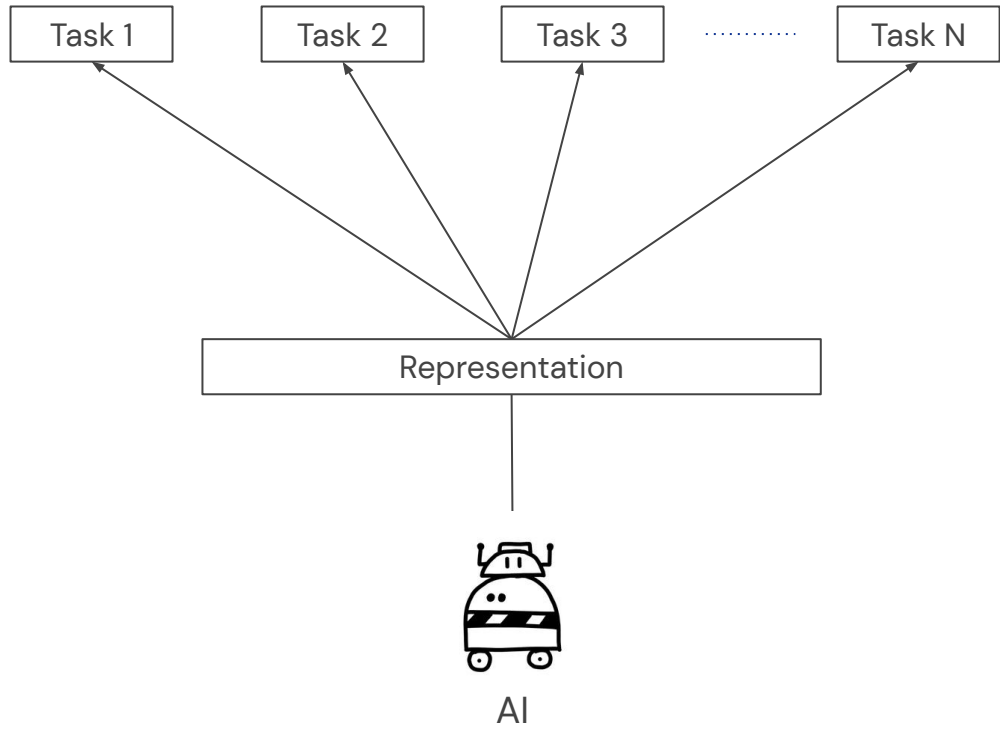


Eviatar Bach / CC BY-SA



Jérémie Barande / Ecole polytechnique Université Paris-Saclay / CC BY-SA 2.0





3

**What  
makes a good  
representation?**





AI



neuroscience



“

**Formal system for making explicit certain entities or types of information, together with a specification of how the system does this.**

– Marr and Nishihara, 1978



Want to learn more?



Representation and Recognition of the Spatial Organization of Three-Dimensional Shapes, Marr & Nishihara, Proc. R. Soc. Lond. 1978



# What is a representation?

Want to learn more?



How Can Deep Learning Advance  
Computational  
Modeling of Sensory Information  
Processing? Thompson et al,  
arxiv 2018

XXXVII

37

0b100101

- Representational form orthogonal to information content
- Useful abstraction to make different computations more efficient
- Not defined by a single piece of information but rather by the shape of the manifold on which the data lie within the representational space

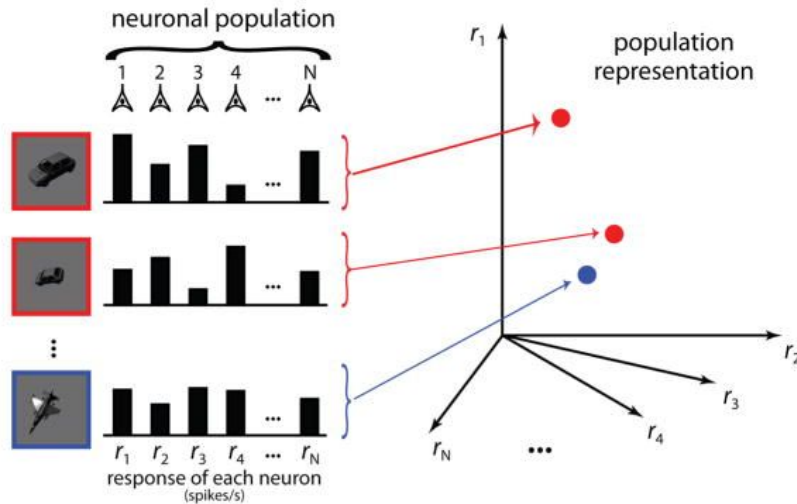


# Untangling representations

Want to learn more?



How Does the Brain Solve Visual Object Recognition?, DiCarlo et al, Neuron 2012

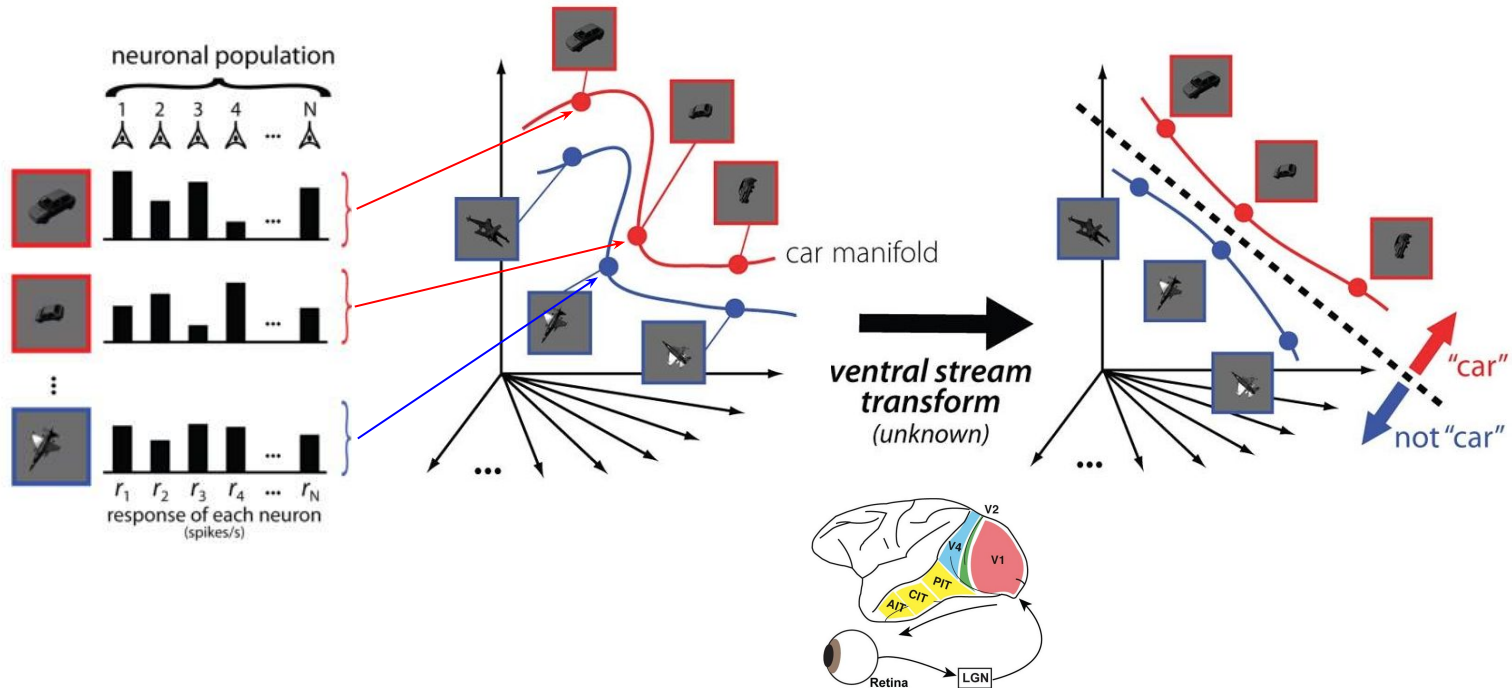


# Untangling representations

Want to learn more?



How Does the Brain Solve Visual Object Recognition?, DiCarlo et al, Neuron 2012





Want to learn more?



Learning Task-State  
Representations, Niv, Nature  
Neuroscience 2019

How does one cross a street?

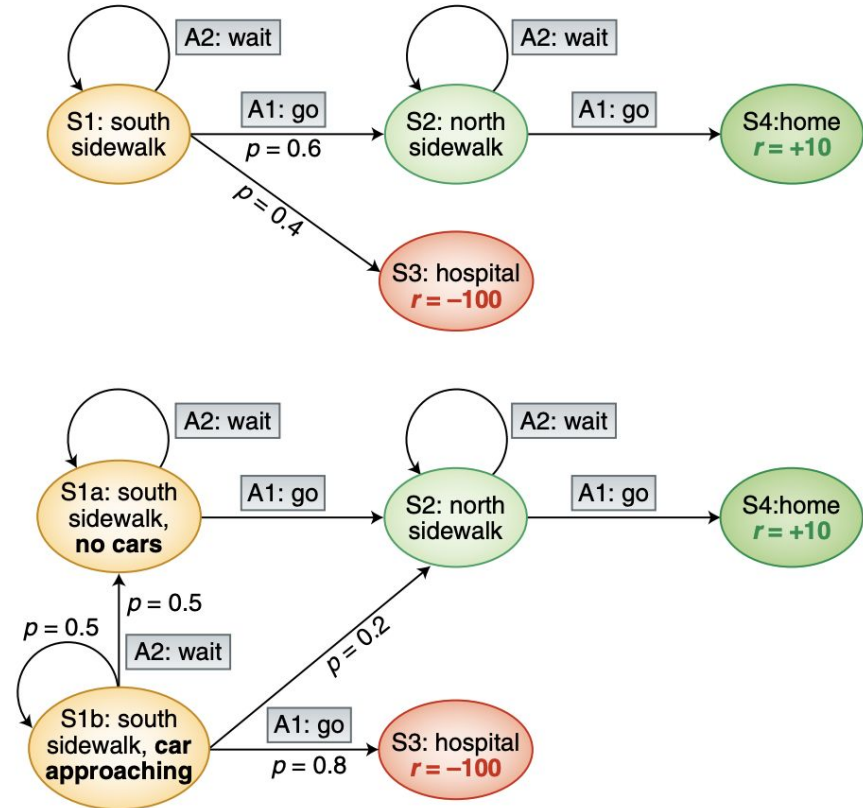


# Alternative representations for the same task

Want to learn more?



Learning Task-State  
Representations, Niv, Nature  
Neuroscience 2019



# Solving tasks requires...

Want to learn more?

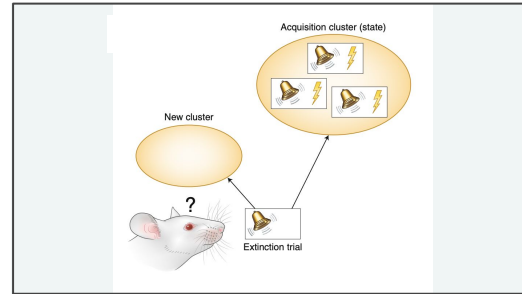


Learning Task-State  
Representations, Niv, Nature  
Neuroscience 2019



## Attention

Representation should support easy attentional attenuation of aspects not relevant to the task.



## Clustering

Experiences should be easily and dynamically clustered together or apart.



## Latent states

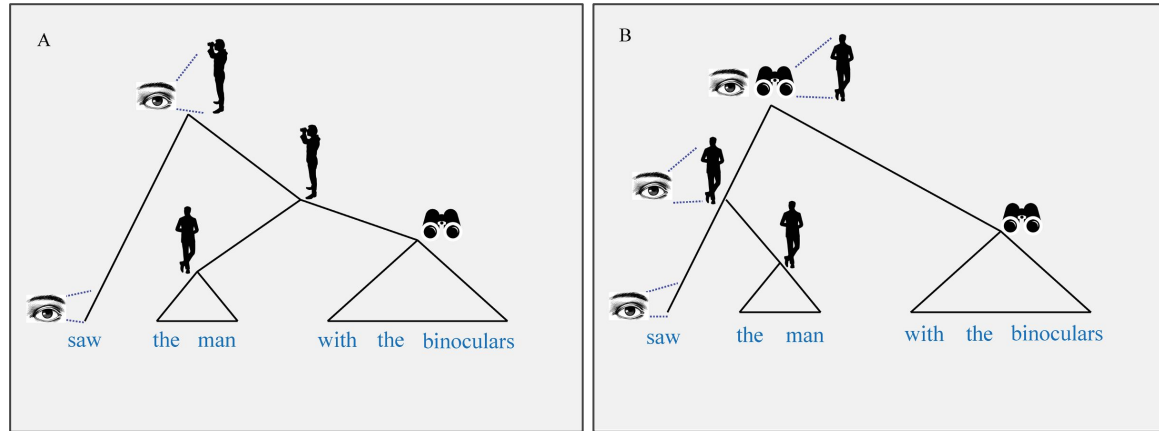
Not all information may be present in perceptual input. Representations should include information about latent aspects of the state too.



# Compositionality

“the meaning of a complex expression is determined by the meanings of its **constituent expressions** and the **rules** used to **combine** them”

Leads to **open-endedness** -- can construct *arbitrarily large number of meaningful complex expressions* from a *finite number of constituent expressions* and *combination rules*.



Bolhuis et al, 2018





physics



AI



neuroscience

- Untangled
- Attention
- Clustering
- Latent information
- Compositionality



# Symmetry transformations

COMMENT



Symmetries feature in the stained-glass ceiling of the Palace of Catalan Music in Barcelona, Spain.

## Why symmetry matters

Mario Livio celebrates the guiding light for modern physics.

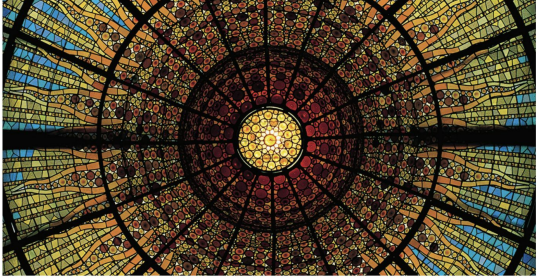
*"It is only slightly overstating the case to say that physics is the study of symmetry."*

– Philip Anderson, 1972



# Symmetry transformations

COMMENT



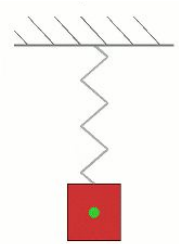
Symmetries feature in the stained-glass ceiling of the Palace of Catalan Music in Barcelona, Spain.

## Why symmetry matters

Mario Livio celebrates the guiding light for modern physics.

*“To a physicist, symmetry is a broader concept than the reflective form of butterfly wings... Symmetry represents those **stubborn cores that remain unaltered** even under transformations that could change them”*

– Mario Livio, 2012



Want to learn more?

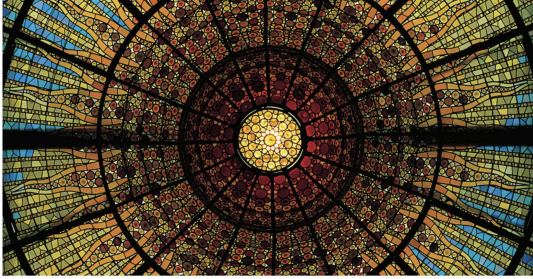


Why symmetry matters,  
Livio, Nature 2012



# Symmetry transformations

COMMENT



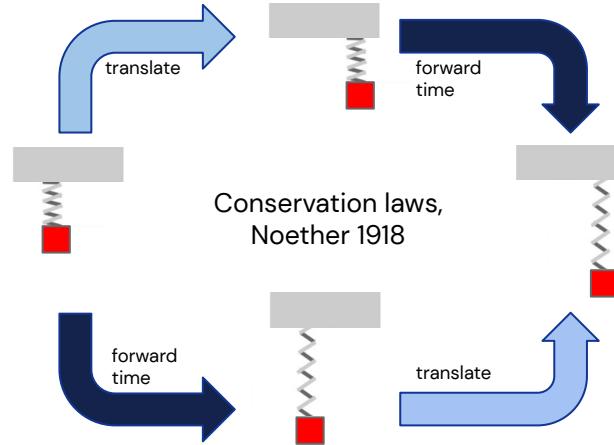
Symmetries feature in the stained-glass ceiling of the Palace of Catalan Music in Barcelona, Spain.

## Why symmetry matters

Mario Livio celebrates the guiding light for modern physics.

*“To a physicist, symmetry is a broader concept than the reflective form of butterfly wings... Symmetry represents those **stubborn cores that remain unaltered** even under transformations that could change them”*

– Livio, 2012



Want to learn more?



Invariante Variationsprobleme,  
Noether, Gesellschaft der  
Wissenschaften zu Göttingen, 1918

Studying symmetries of a system helps:

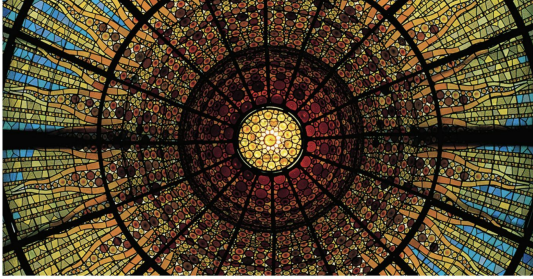
- Unify existing theories (e.g. electromagnetism)
- Categorise physical objects (e.g. elementary particles)
- Discover new physical objects (e.g. particle  $\Omega^-$  - predicted in 1962, discovered in 1964)





# Symmetry transformations

COMMENT



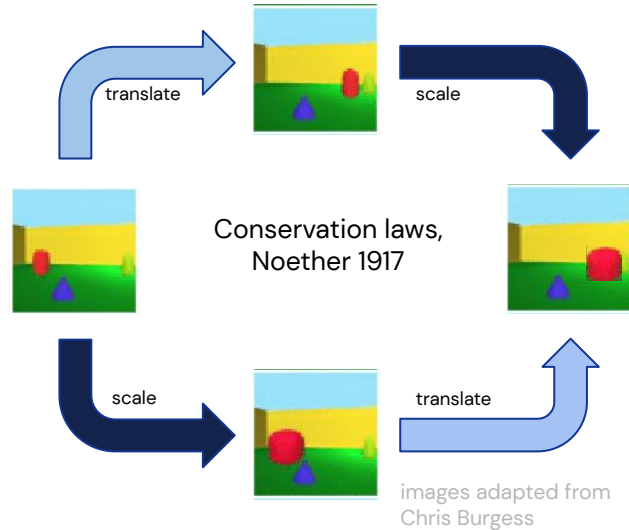
Symmetries feature in the stained-glass ceiling of the Palace of Catalan Music in Barcelona, Spain.

## Why symmetry matters

Mario Livio celebrates the guiding light for modern physics.

*“To a physicist, symmetry is a broader concept than the reflective form of butterfly wings... Symmetry represents those **stubborn cores that remain unaltered** even under transformations that could change them”*

– Livio, 2012



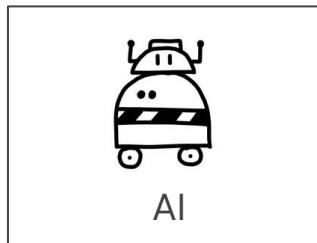
Studying symmetries of a system helps:

- Unify existing theories (e.g. electromagnetism)
- Categorise physical objects (e.g. elementary particles)
- Discover new physical objects (e.g. particle  $\Omega^-$  - predicted in 1962, discovered in 1964)





physics



AI



neuroscience

→ Symmetries

→ Untangled

→ Attention

→ Clustering

→ Latent information

→ Compositionality



# Information bottleneck

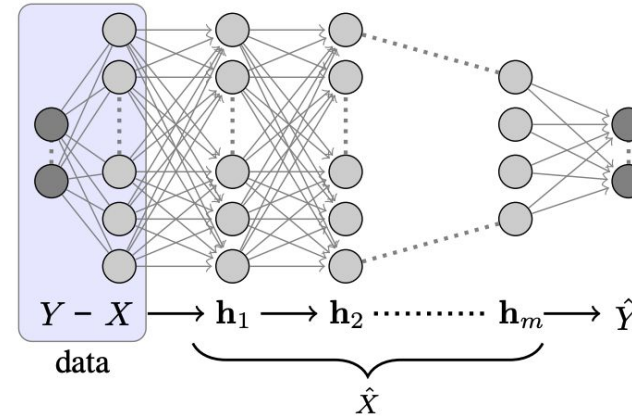
Want to learn more?



Deep Learning and the  
Information Bottleneck  
Principle, Tishby & Zaslavsky,  
IEEE ITW 2015

Goal of supervised learning:

“find a **maximally compressed** mapping of the input variable that **preserves** as much as possible the **information** on the output variable.”

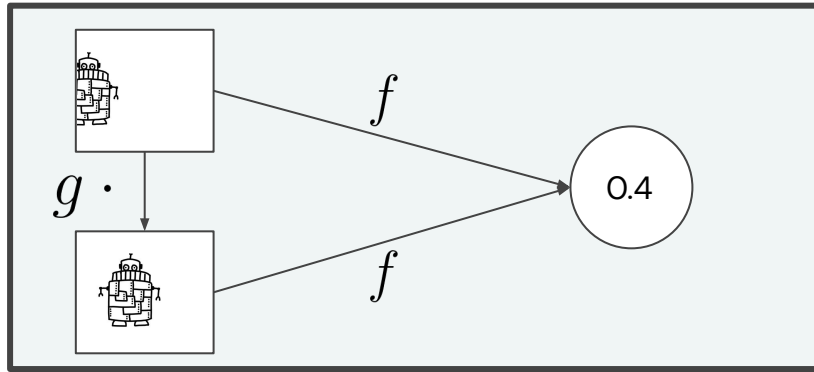


$$I(Y; X) \geq I(Y; \mathbf{h}_j) \geq I(Y; \mathbf{h}_i) \geq I(Y; \hat{Y})$$

**Data processing inequality** (Shannon, 1948) --  
post-processing cannot increase information.



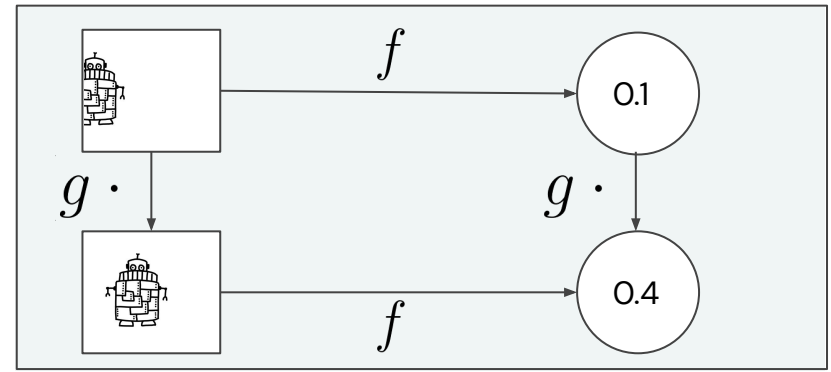
# Invariance vs equivariance



## Invariance

- representation remains unchanged when a certain type of transformation is applied to the input

$$f(g \cdot x) = f(x)$$



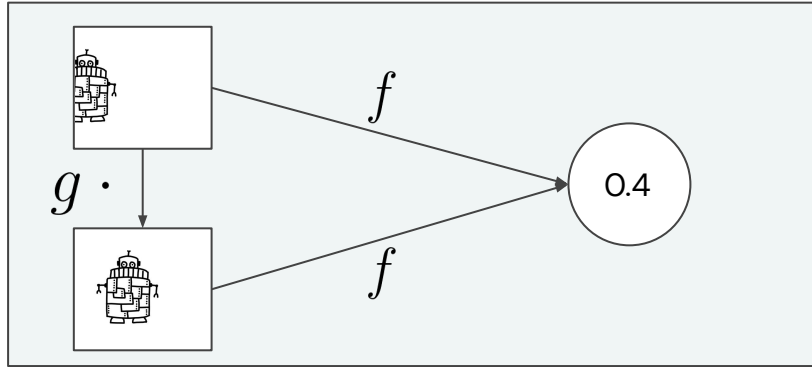
## Equivariance

- representation reflects the transformation applied to the input

$$f(g \cdot x) = g \cdot f(x)$$



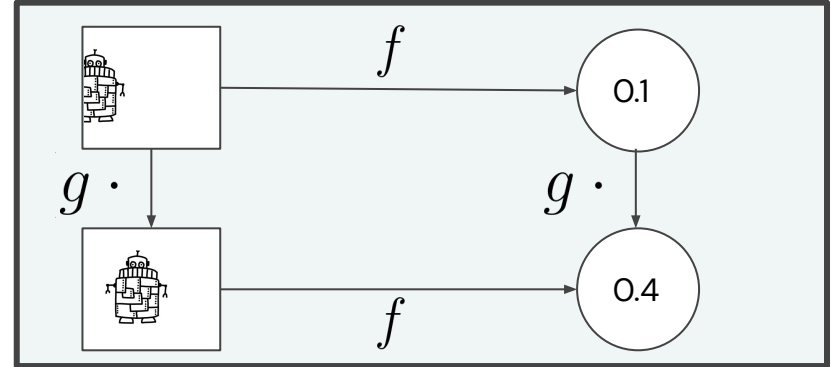
# Invariance vs equivariance



## Invariance

- representation remains unchanged when a certain type of transformation is applied to the input

$$f(g \cdot x) = f(x)$$



## Equivariance

- representation reflects the transformation applied to the input

$$f(g \cdot x) = g \cdot f(x)$$

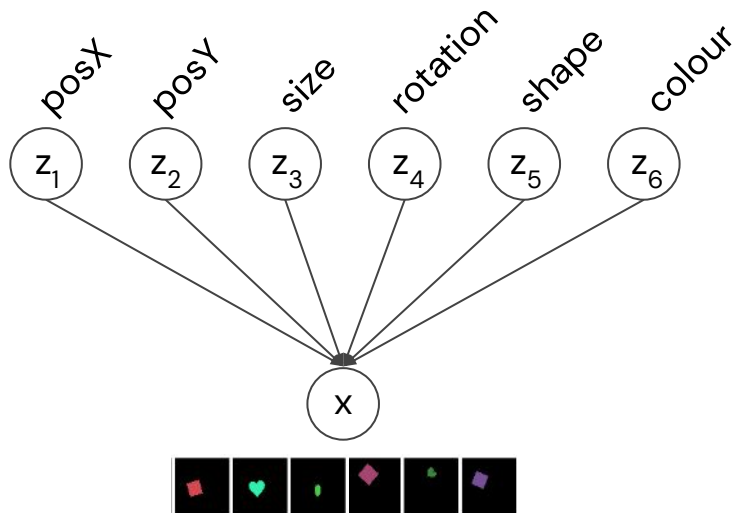


# Disentangled representation learning

Want to learn more?

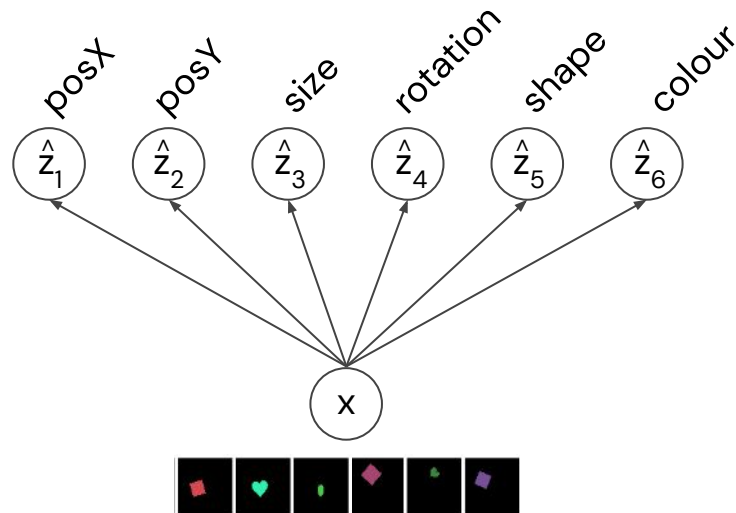


Deep Learning of  
Representations: Looking  
Forward, Bengio, SLSP 2013



$$p(z) = \prod_i p(z_i)$$

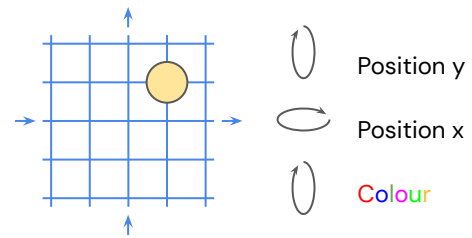
Generative process



$$p(x, z) = p(x, \hat{z})$$

Inference process

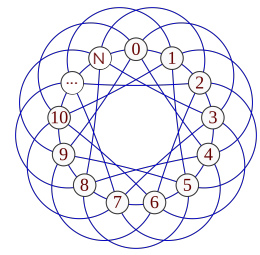




$$G_x = C_N$$

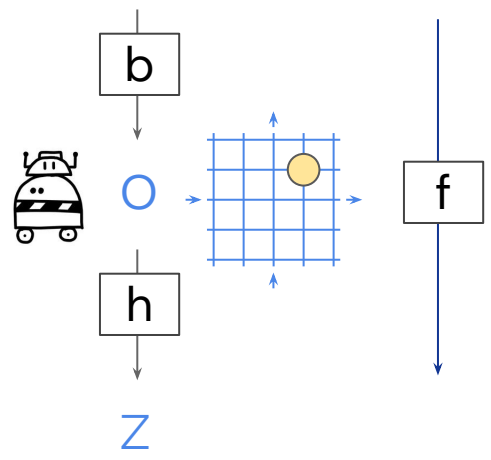
$$G_y = C_N$$

$$G_c = C_N$$

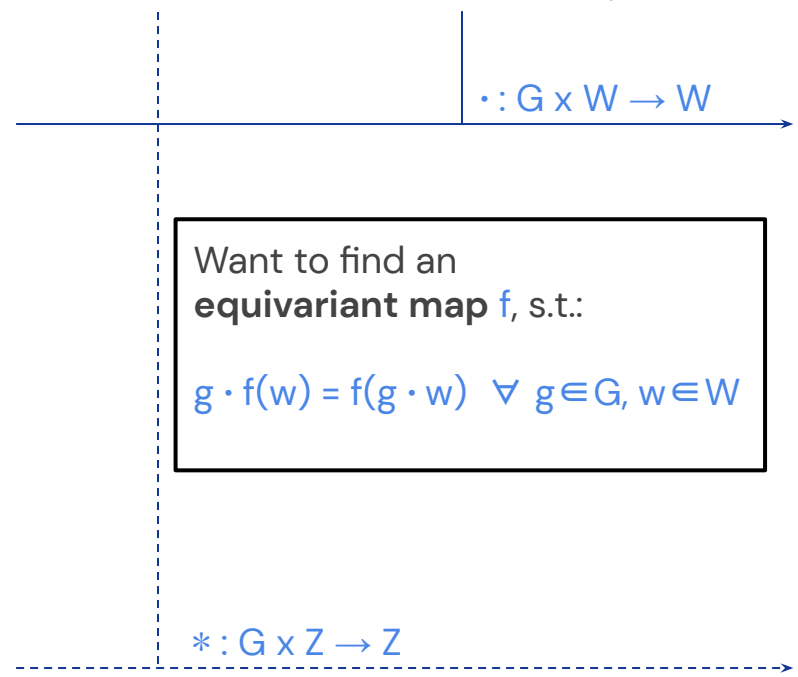


Symmetry group  $G = G_x \times G_y \times G_c$

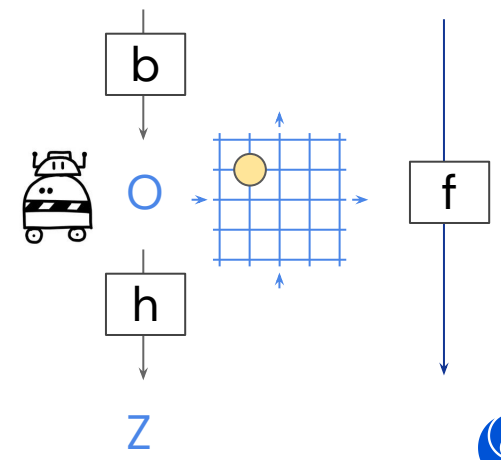
$W$   
 (x=5, y=6, c=yellow)



$[z_x = 0.9, z_y = 0.3, z_c = 0.1]$



$W$   
 (x=3, y=6, c=yellow)



$[z_x = 0.3, z_y = 0.3, z_c = 0.1]$



# 4

## Evaluating the merit of a representation

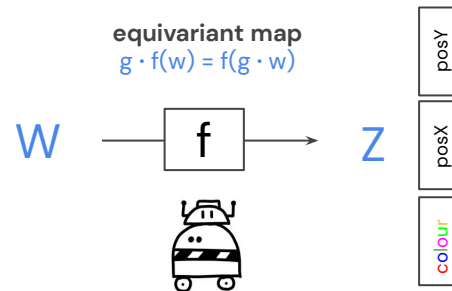
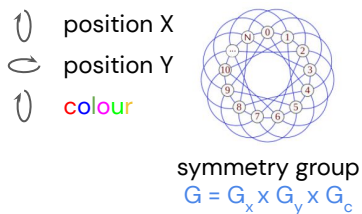




# Evaluating representations (example)

To support efficient solving of numerous diverse held-out tasks, representation should be/support:

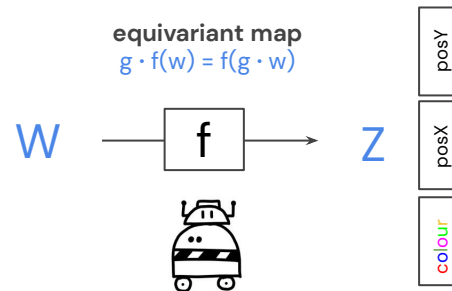
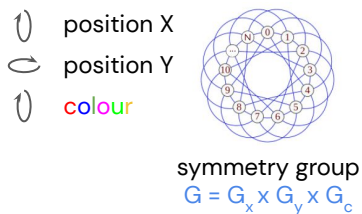
- Symmetries
- Untangled
- Attention
- Clustering
- Compositionality



# Evaluating representations (example)

To support efficient solving of numerous diverse held-out tasks, representation should be/support:

- Symmetries
- Untangled
- Attention
- Clustering
- Compositionality

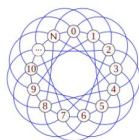


# Evaluating representations (example)

To support efficient solving of numerous diverse held-out tasks, representation should be/support:

- Symmetries
- Untangled
- Attention
- Clustering
- Compositionality

↻ position X  
↻ position Y  
↻ colour

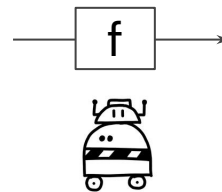


symmetry group  
 $G = G_x \times G_y \times G_c$

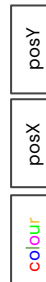
group operator  
 $(g_1, h_1) \bullet (g_2, h_2) = (g_1 \circ g_2, h_1 \circ h_2)$

equivariant map  
 $g \cdot f(w) = f(g \cdot w)$

W



Z

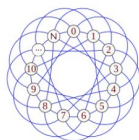


# Evaluating representations (example)

To support efficient solving of numerous diverse held-out tasks, representation should be/support:

- Symmetries
- Untangled
- Attention
- Clustering
- Compositionality

↻ position X  
↻ position Y  
↻ colour



symmetry group  
 $G = G_x \times G_y \times G_c$

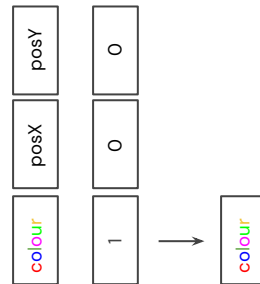
group operator  
 $(g_1, h_1) \bullet (g_2, h_2) = (g_1 \circ g_2, h_1 \circ h_2)$

equivariant map  
 $g \cdot f(w) = f(g \cdot w)$

W



Z



attention  
(e.g. binary mask)

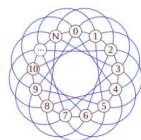


# Evaluating representations (example)

To support efficient solving of numerous diverse held-out tasks, representation should be/support:

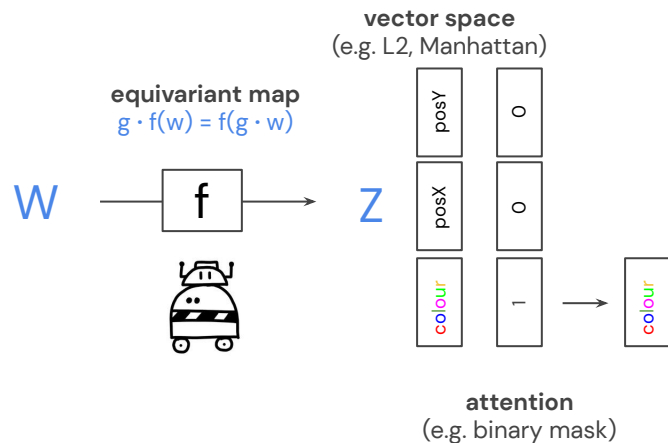
- Symmetries
- Untangled
- Attention
- Clustering
- Compositionality

↻ position X  
↻ position Y  
↻ colour



symmetry group  
 $G = G_x \times G_y \times G_c$

group operator  
 $(g_1, h_1) \bullet (g_2, h_2) = (g_1 \circ g_2, h_1 \circ h_2)$



# Evaluating representations (example)

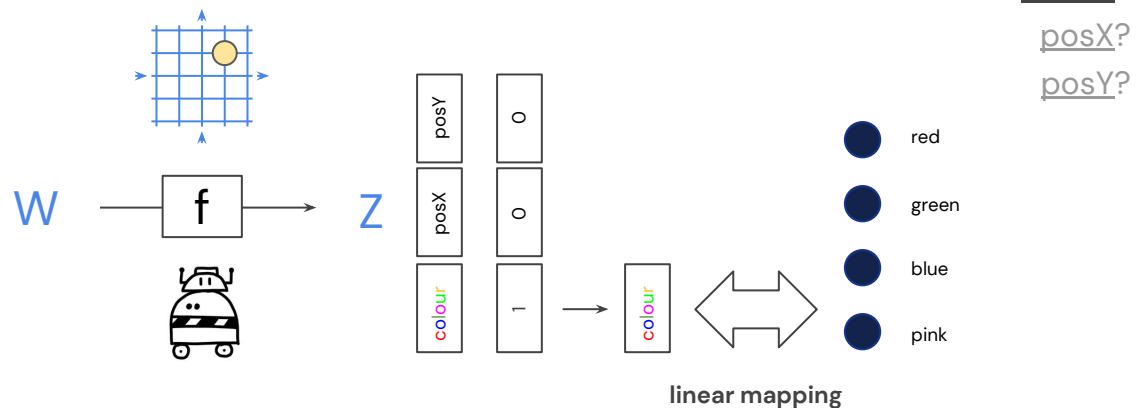
Want to learn more?



Deep Symmetry Networks,  
Gens and Domingos, NeurIPS  
2014

Such a representation should help with:

- Data efficiency
- Robustness
- Generalisation
- Transfer
- "Common sense"



# Evaluating representations (example)

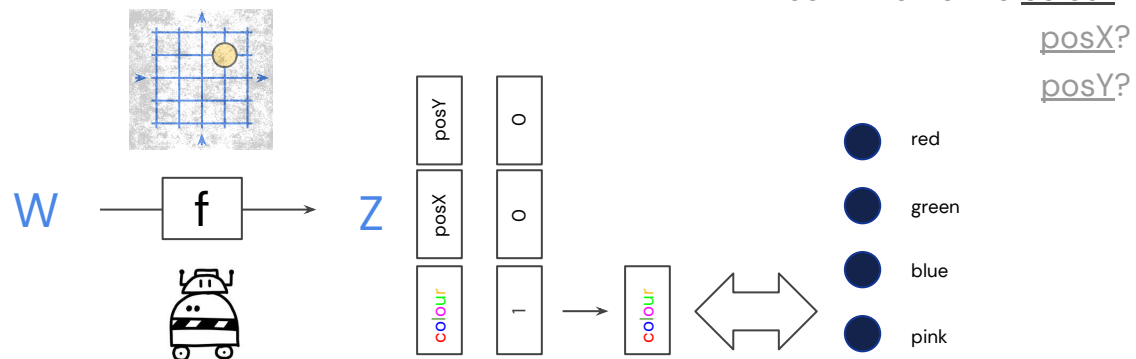
Want to learn more?



Achieving Robustness in the Wild via Adversarial Mixing with Disentangled Representations, Goyal et al, CoRR 2019

Such a representation should help with:

- Data efficiency
- Robustness
- Generalisation
- Transfer
- "Common sense"



# Evaluating representations (example)

Want to learn more?



Towards Robust Image Classification Using Sequential Attention Models, Zoran et al, CVPR 2020

Such a representation should help with:

- Data efficiency
- Robustness
- Generalisation
- Transfer
- "Common sense"

source: wallet  
target: beaver



top-1: wallet

**ResNet-152**  
250 PGD steps



top-1: beaver

**S3TA-8 (Ours)**  
250 PGD steps



top-1: wallet





# Evaluating representations (example)

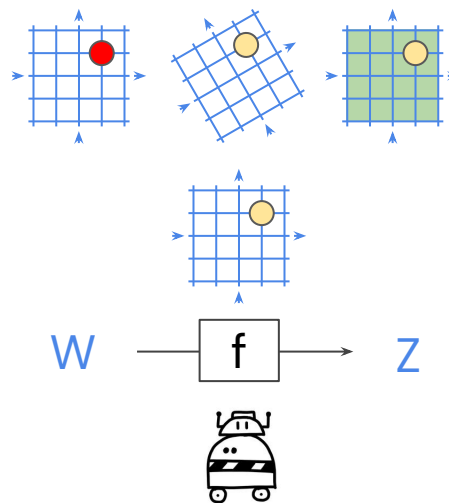
Want to learn more?



DARLA: Improving Zero-Shot Transfer in Reinforcement Learning, Higgins, Pal et al, ICML 2017

Such a representation should help with:

- Data efficiency
- Robustness
- Generalisation
- Transfer
- "Common sense"



Task: go to bottom left corner?



# Evaluating representations (example)

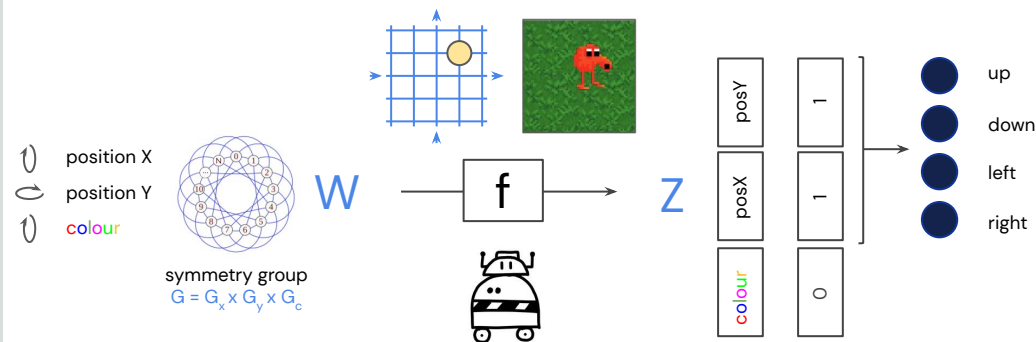
Want to learn more?



Life-Long Disentangled Representation Learning with Cross-Domain Latent Homologies, Achille et al, NeurIPS 2018

Such a representation should help with:

- Data efficiency
- Robustness
- Generalisation
- Transfer
- "Common sense"



# Evaluating representations (example)

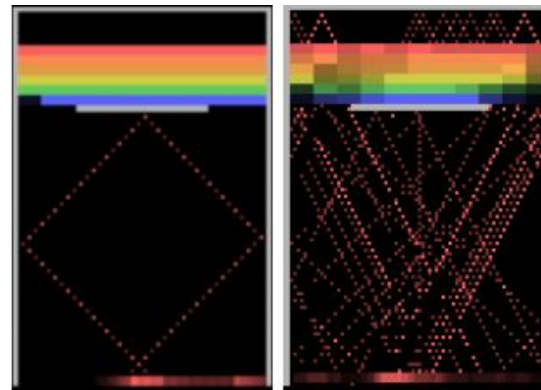
Want to learn more?



Schema Networks: Zero-shot Transfer with a Generative Causal Model of Intuitive Physics, Kansky et al, ICML 2017

Such a representation should help with:

- Data efficiency
- Robustness
- Generalisation
- Transfer
- “Common sense”



(a) A3C

(b) Schema Networks

	Standard Breakout	Offset Paddle	Middle Wall	Random Target	Juggling
A3C Image Only	N/A	$0.60 \pm 20.05$	$9.55 \pm 17.44$	$6.83 \pm 5.02$	$-39.35 \pm 14.57$
A3C Image + Entities	N/A	$11.10 \pm 17.44$	$8.00 \pm 14.61$	$6.88 \pm 6.19$	$-17.52 \pm 17.39$
Schema Networks	$36.33 \pm 6.17$	$41.42 \pm 6.29$	$35.22 \pm 12.23$	$21.38 \pm 5.02$	$-0.11 \pm 0.34$



# Evaluating representations (example)

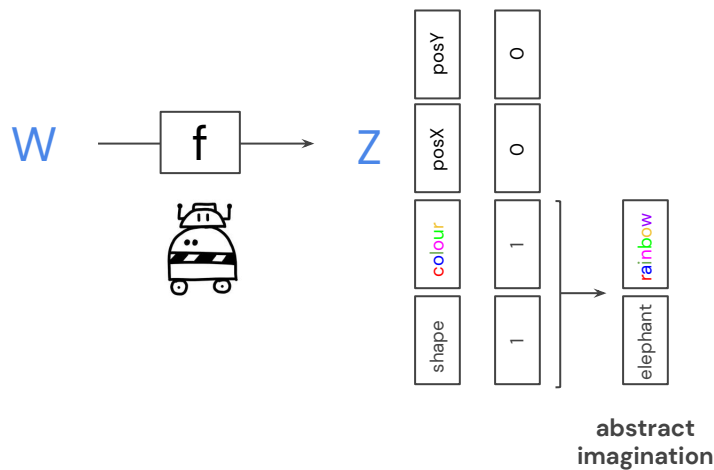
Want to learn more?



SCAN: Learning Hierarchical  
Compositional Visual Concepts,  
Higgins et al, ICLR 2018

Such a representation should help with:

- Data efficiency
- Robustness
- Generalisation
- Transfer
- “Common sense”



# Evaluating representations (example)

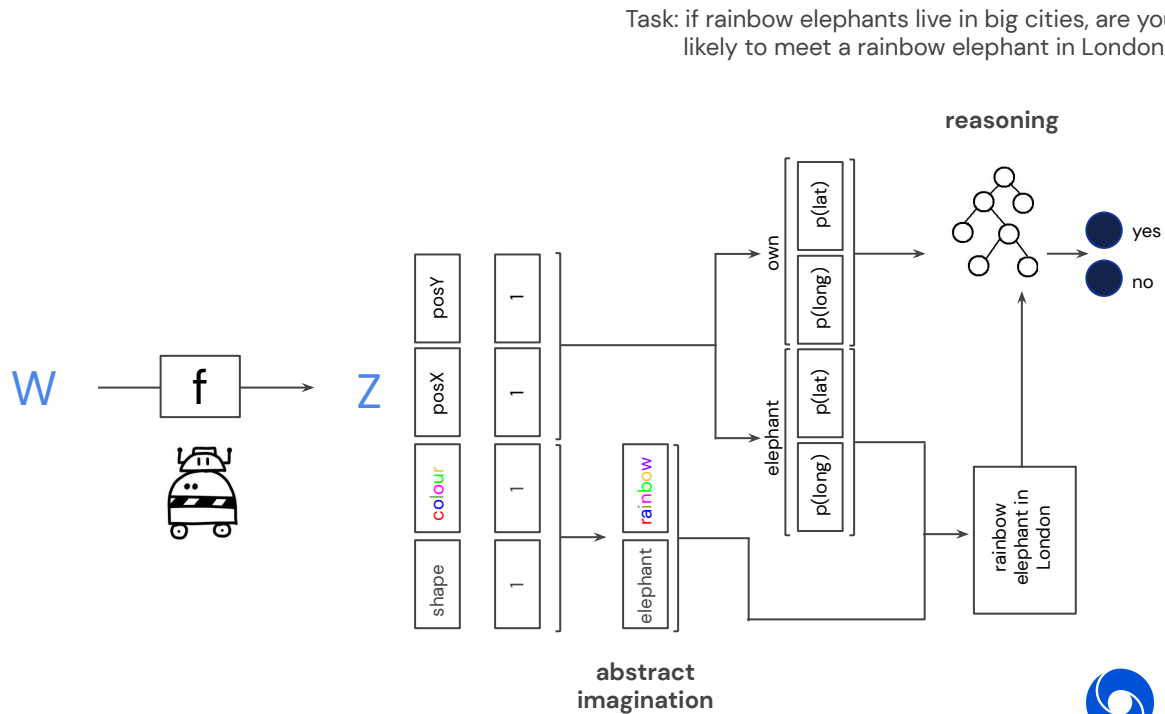
Want to learn more?



Human-Level Concept Learning  
Through Probabilistic Program  
Induction, Lake et al, Science  
2015

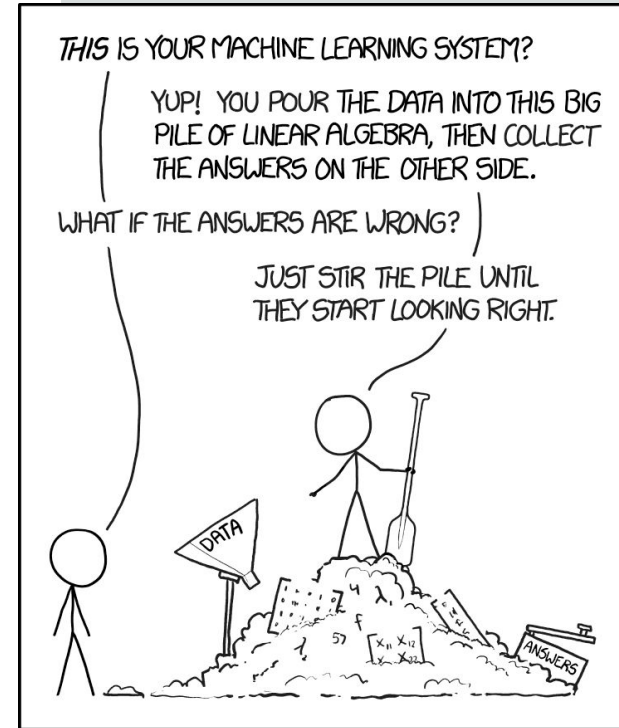
Such a representation should help with:

- Data efficiency
- Robustness
- Generalisation
- Transfer
- “Common sense”



# Is all machine learning ultimately about representation learning?

- Crucial early role of representations for ML (hand crafted feature engineering)
- Success of supervised deep learning on single tasks may be attributed to good implicit representation learning (information bottleneck principle)
- Lack of understanding and control over the nature of learnt representations may be behind current problems with deep learning
- Recent advances in deep learning may be attributed to learning better representations (e.g. Devlin et al, 2018; Chen et al, 2019; Lyle et al, 2019)
- Further advancement of deep learning may benefit from explicit/unsupervised representation learning
- Advancing explicit/unsupervised representation learning may benefit from interdisciplinary insights



[xkcd.com/1838](https://xkcd.com/1838)



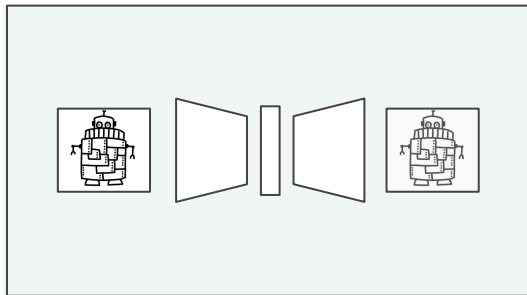
DeepMind

5

# Representation learning techniques

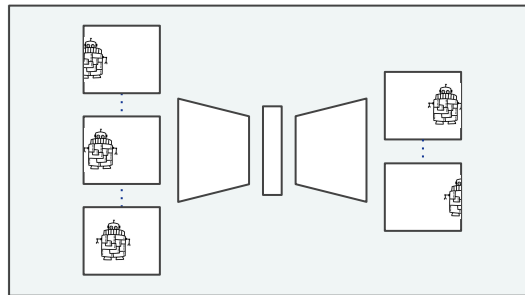


# Representation learning with deep neural networks



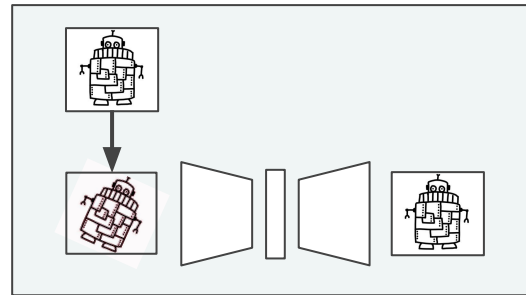
## Generative modeling

Learn the data distribution using generative modeling, often through reconstructions.



## Contrastive losses

Use classification losses to learn representations that preserve temporal or spatial data consistency.



## Self-supervision

Exploit knowledge of data to design learning tasks which lead to useful representations.





# Downstream tasks for representation learning

## Semi supervised learning

Use the learned representations for classification.

**Aim:** data efficiency, generalization.  
standard benchmark: Imagenet.

## Reinforcement learning

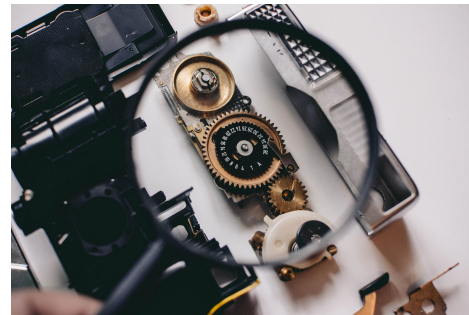
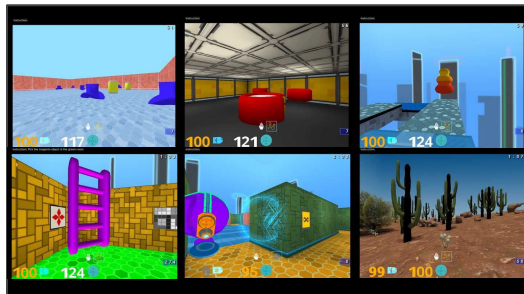
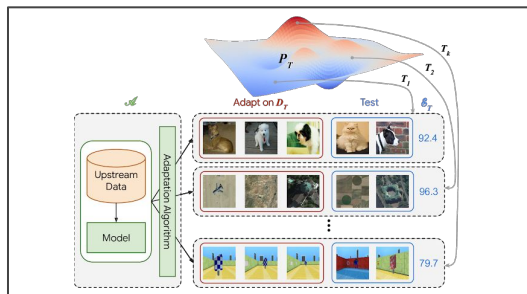
Use the learned representations for model based RL or model free RL.

**Aim:** data efficiency, transfer.

## Model analysis

Use the learned representation to analyse what the model learns.

**Aim:** Interpretable models.



## Keep in mind that we want....

- Discrete and continuous representations
- Online learning (representations adapt with experience)
- Consistency and temporal abstractions
- Data efficiency
- Downstream tasks

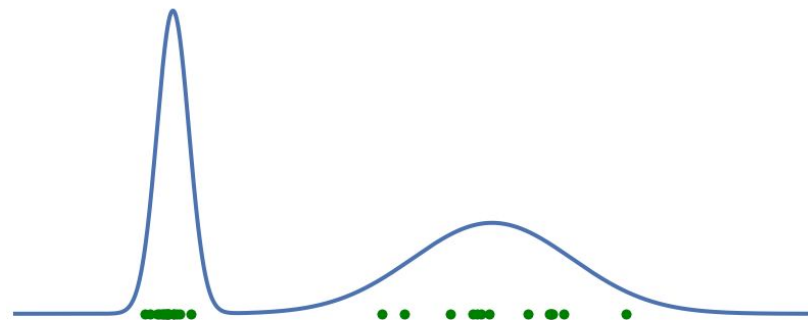


# Generative modeling



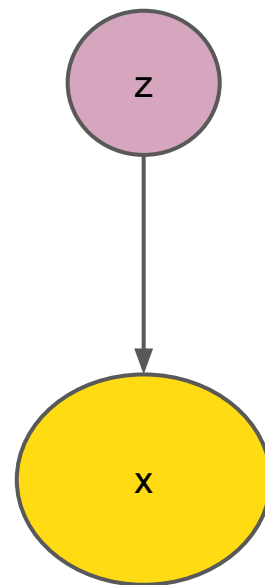
# Generative modelling for representation learning

- model the underlying data distribution
- unsupervised learning (task agnostic)
- Intuition: the most efficient way to model a distribution is to extract common patterns (representations)



# Latent variable models

Model the data generating process as a mapping from a low dimensional unknown (latent) space to the data distribution.

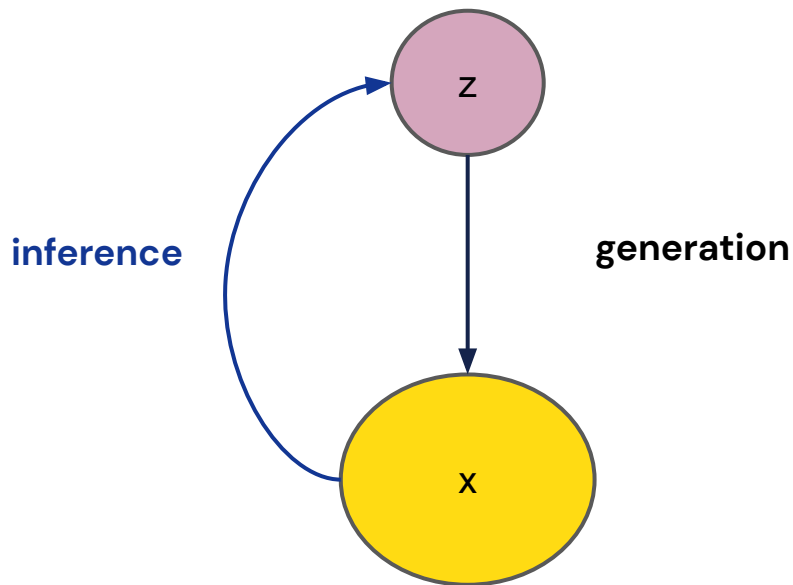


# Inference in latent variable models

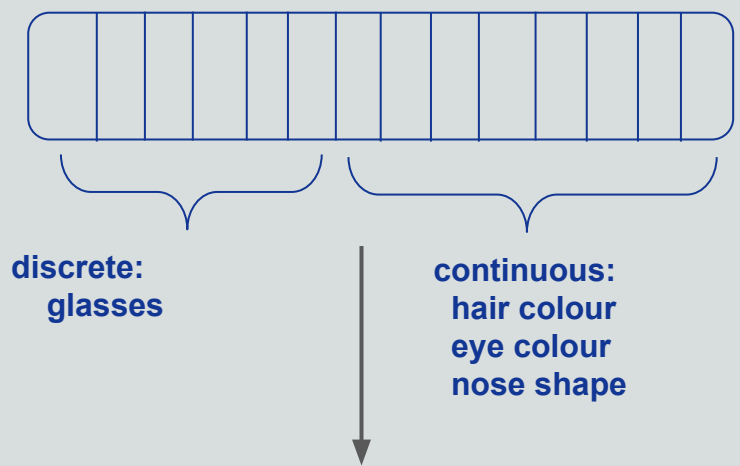
Inference: Find  $p(z|x)$

Intuition: Find the underlying factors which generated the data (with uncertainty estimates).

Finding  $p(z|x)$  is often intractable, and we thus have to resort to approximations.



# Generation



**discrete:**  
glasses

**continuous:**  
hair colour  
eye colour  
nose shape



# Inference



# Variational autoencoders

Maximum likelihood

$$\mathbb{E}_{p^*(\mathbf{x})} [\log p_{\theta}(\mathbf{x})]$$

Latent variable model

$$\log p_{\theta}(\mathbf{x}) = \log \int p_{\theta}(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$$





# Variational autoencoders

Want to learn more?



Auto-Encoding Variational  
Bayes, Kingma et al., ICLR 2017

Lower bound on maximum likelihood objective (ELBO):

$$\log p_{\theta}(\mathbf{x}) \geq \underbrace{\mathbb{E}_{q_{\eta}(\mathbf{z}|\mathbf{x})} \log p_{\theta}(\mathbf{x}|\mathbf{z})}_{\text{reconstruct}} - \underbrace{\text{KL}(q_{\eta}(\mathbf{z}|\mathbf{x}) || p(\mathbf{z}))}_{\text{stay close to prior}}$$

Approximate posterior

$$q_{\eta}(\mathbf{z}|\mathbf{x})$$



# VAEs - the role of the prior

Want to learn more?



Auto-Encoding Variational  
Bayes, Kingma et al., ICLR 2017

The KL term regularises the approximate posterior to the prior.

Use the prior to specify properties we would like the posterior to have, such as disentanglement.

$$KL[q_{\eta}(\mathbf{z}|\mathbf{x})]||| [p(\mathbf{z})]$$



# VAEs and neural networks

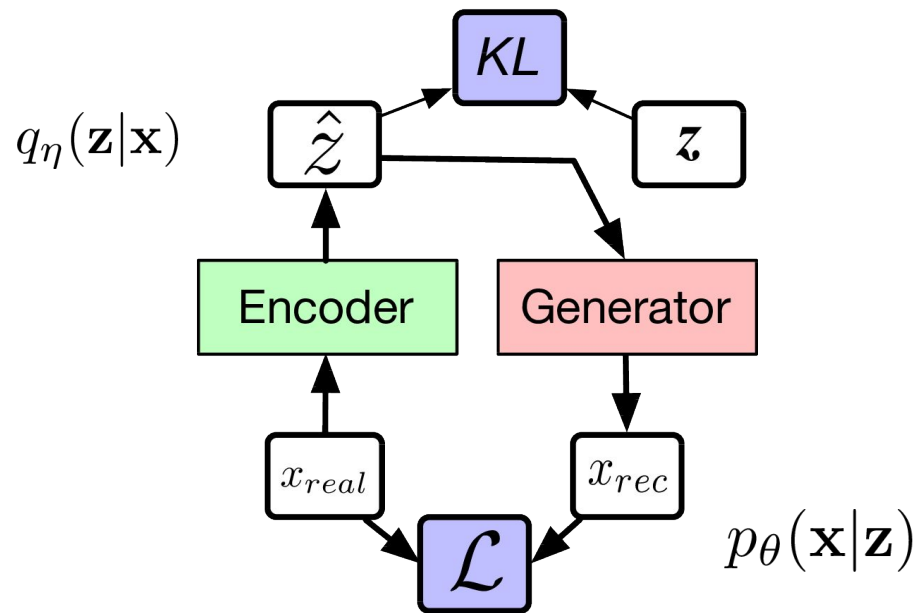
$$q_{\eta}(\mathbf{z}|\mathbf{x})$$

$$p_{\theta}(\mathbf{x}|\mathbf{z})$$

Both the inference and generation model are deep neural networks.

Want to learn more?

Auto-Encoding Variational  
Bayes, Kingma et al., ICLR 2017





$$\mathbb{E}_{q_{\eta}(\mathbf{z}|\mathbf{x})} \log p_{\theta}(\mathbf{x}|\mathbf{z}) - \beta \text{KL}(q_{\eta}(\mathbf{z}|\mathbf{x}) || p(\mathbf{z}))$$

Change the weight of the KL term to encourage disentangled representations.



# beta-VAE

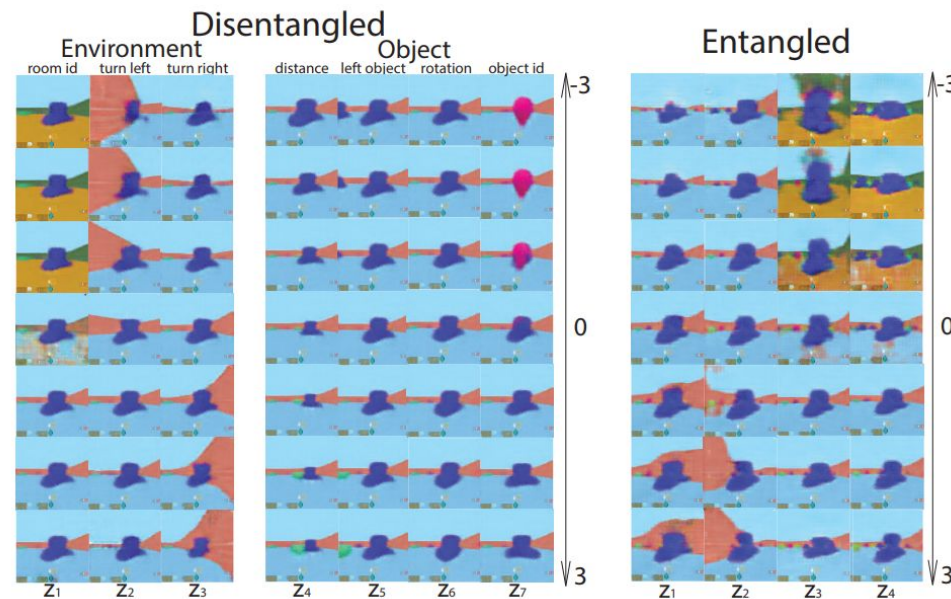
Learns disentangled continuous representations encoding semantic information:

location, object, distance to object, rotations.

Want to learn more?



beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework, Higgins et al., ICLR 2017



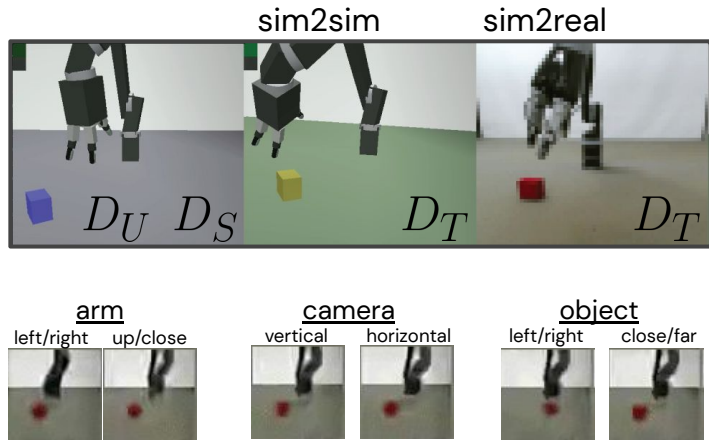
# beta-VAE in reinforcement learning

Integrating beta-VAEs into reinforcement learning agents improves generalization and transfer.

Want to learn more?



DARLA: Improving Zero-Shot Transfer in Reinforcement Learning  
Higgins et al., ICML 2017



VISION TYPE	JACO (A3C)	
	SIM2SIM	SIM2REAL
BASELINE AGENT UNREAL	97.64 ± 9.02	94.56 ± 3.55
DARLA <sub>FT</sub>	86.59 ± 5.53	99.25 ± 2.3
DARLA <sub>ENT</sub>	84.77 ± 4.42	59.99 ± 15.05
DARLA <sub>DAE</sub>	85.15 ± 7.43	100.72 ± 4.7
<b>DARLA</b>	<b>100.85 ± 2.92</b>	<b>108.2 ± 5.97</b>



# Sequential VAEs - ConvDraw

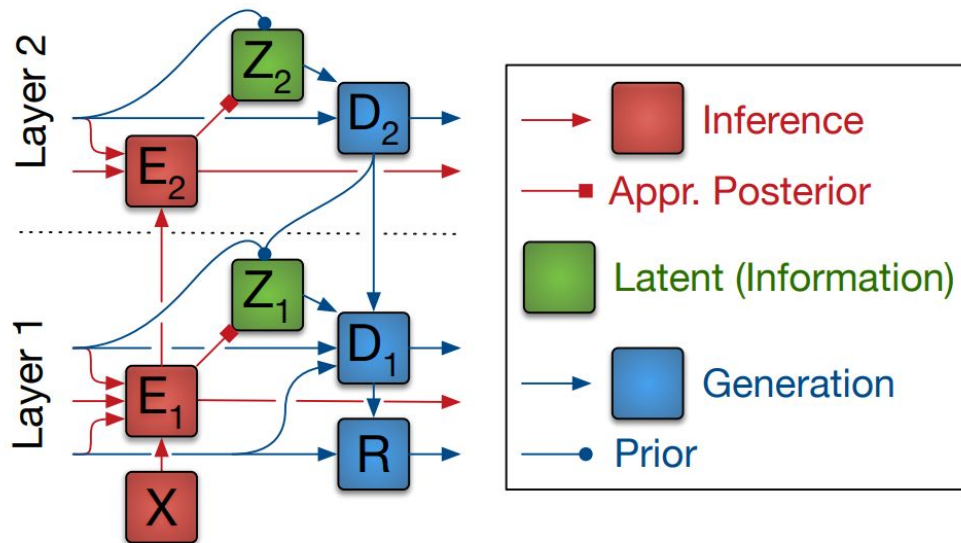
Want to learn more?



Towards Conceptual  
Compression  
Gregor et al., NIPS 2016

→ VAE (reconstruction and KL loss)

→ Recurrent component



# Conv-Draw

- Recurrence helps: iteratively refine and add details.
- Inference: powerful autoregressive posteriors.
- Latents: spatial and temporal.

Want to learn more?



Towards Conceptual  
Compression  
Gregor et al., NIPS 2016





# Layered models - Monet

Want to learn more?



MONet: Unsupervised  
Scene Decomposition and  
representation  
Burgess et al., arxiv 2019

- VAE and segmentation network
- VAE input learned using attention
- Compositional: adding masks leads to final image

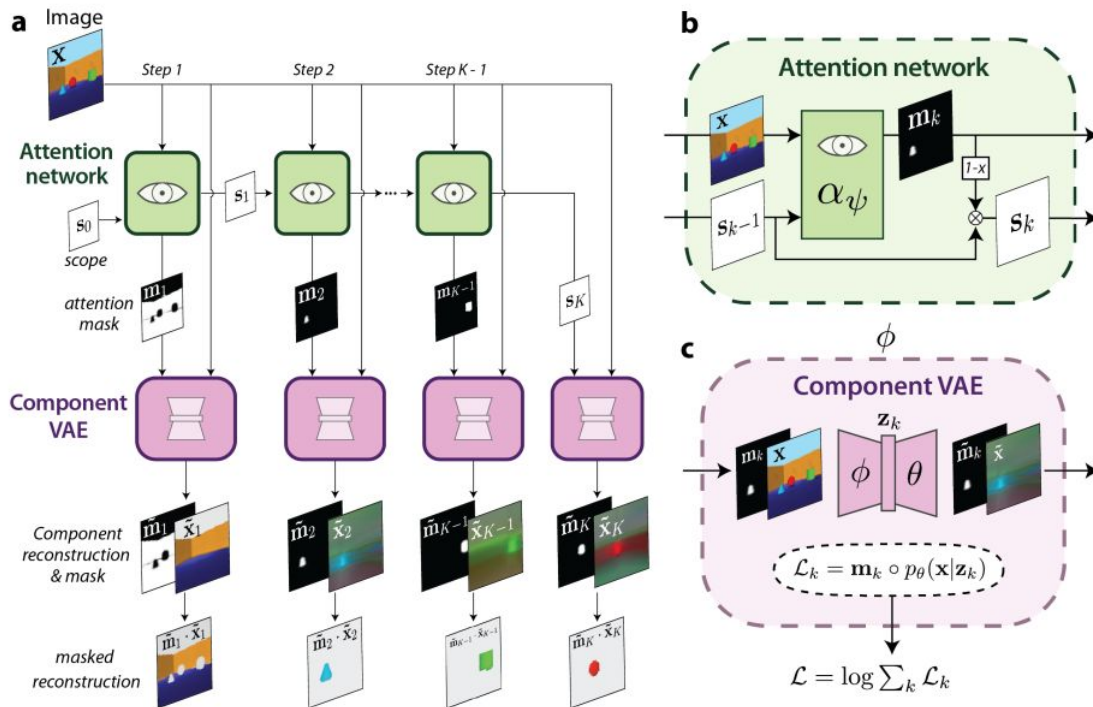


Figure from Burgess et al. (2019)

# MONET

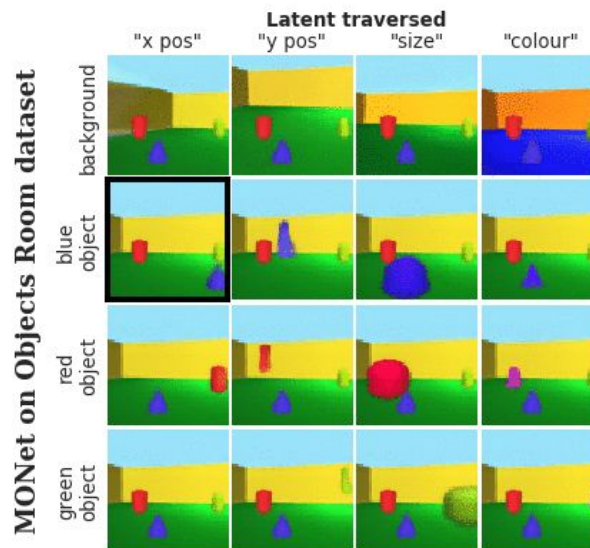
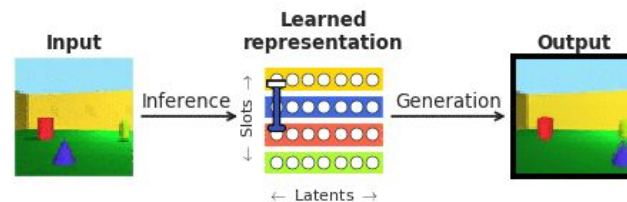
Using attention in a multi level process leads to a generative model which learns concepts (objects) unsupervised.

Latent traversals show that Monet learns to encode the position of an object into a single latent.

Want to learn more?



MONet: Unsupervised  
Scene Decomposition and  
representation  
Burgess et al., arxiv 2019



# MONET in reinforcement learning

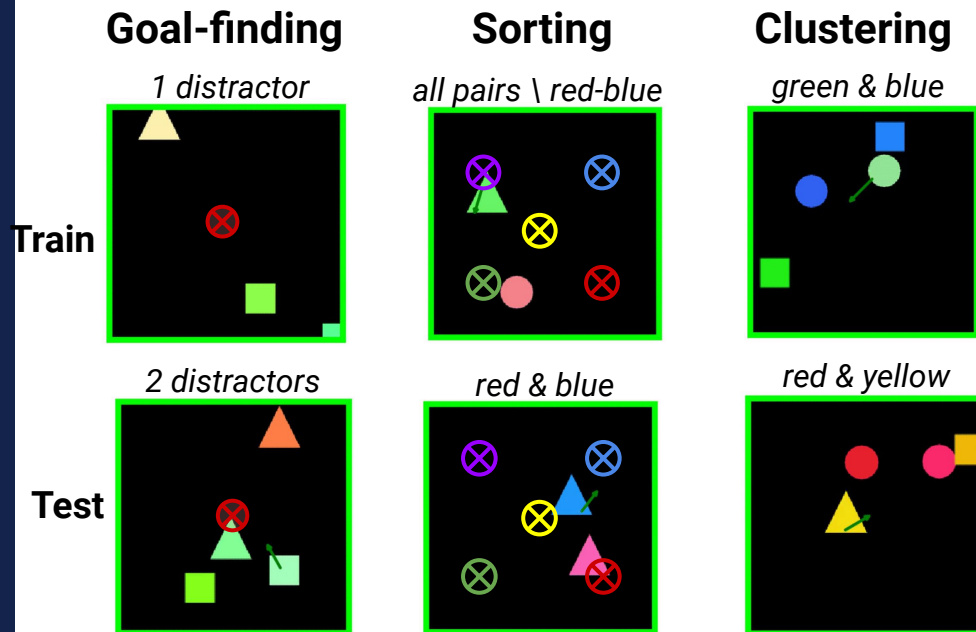
After unsupervised exploration phase, agent to learn tasks quickly.

Given scene representation, transition model, and exploration policy, agent must only learn a reward function, then can do model-based search.

Want to learn more?



MONet: Unsupervised  
Scene Decomposition and  
representation  
Burgess et al., arxiv 2019



# Generative Query networks (GQN)

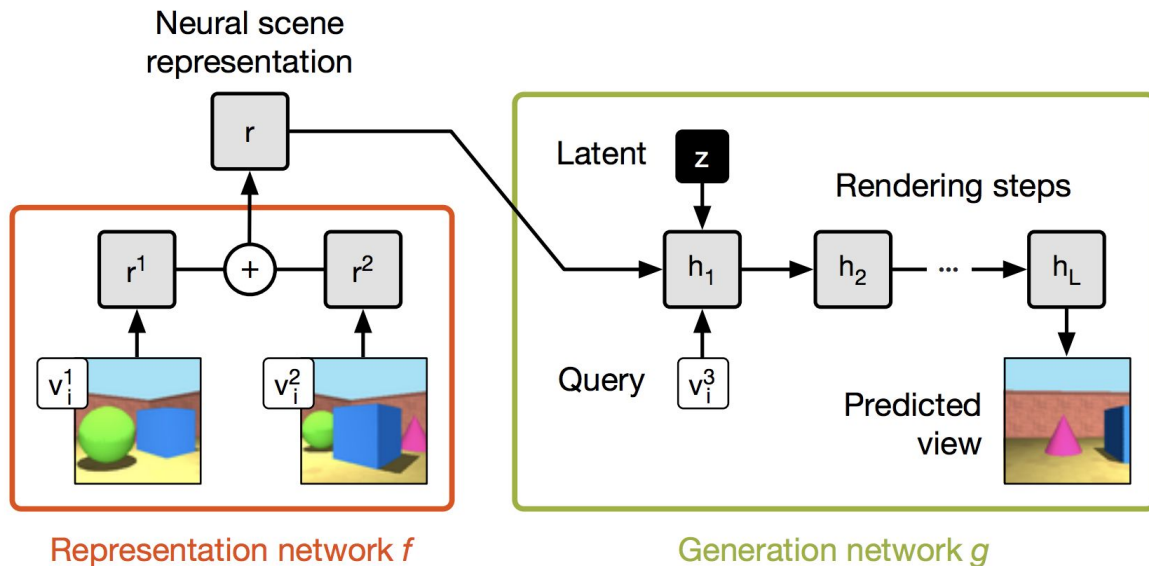
Want to learn more?



Eslami et al, Neural scene representation and rendering, Science (2018)

→ Learn a representation by providing examples angle scene pairs

→ Use learned representation to condition a (recurrent) generative model to generate how the scene looks like from a different angle



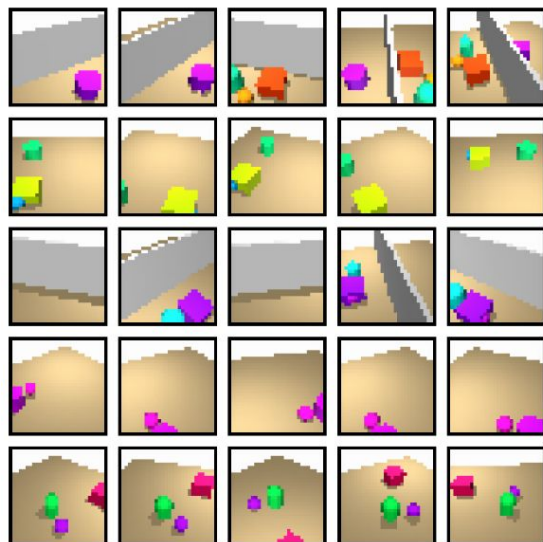
Slides thanks to Ali Eslami.

# GQN - accurate generation

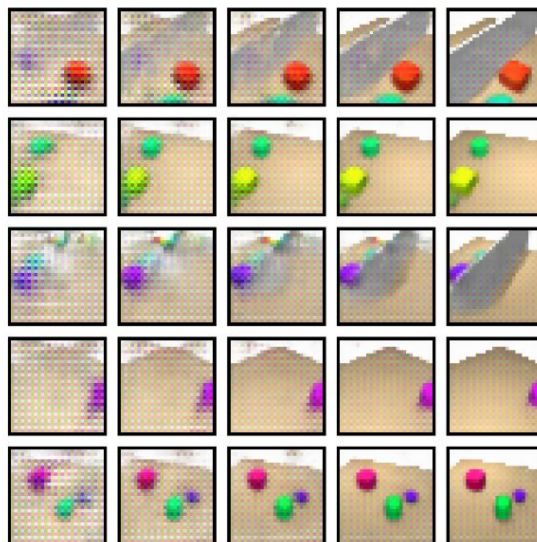
Want to learn more?



Eslami et al, Neural scene representation and rendering, Science (2018)



Observations



Generation steps



Pred

Truth

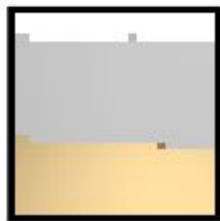


# GQN - capturing uncertainty

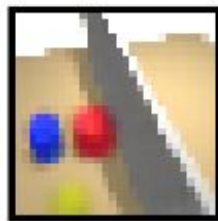
Want to learn more?



Eslami et al, Neural scene representation and rendering, Science (2018)



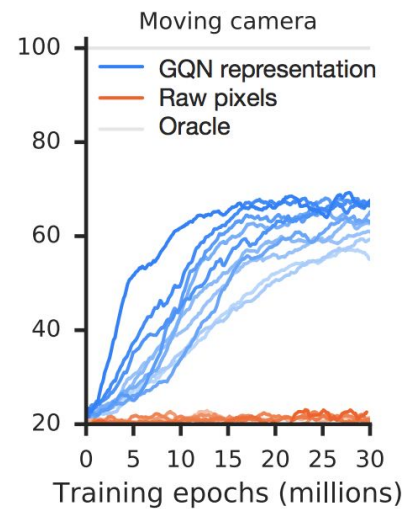
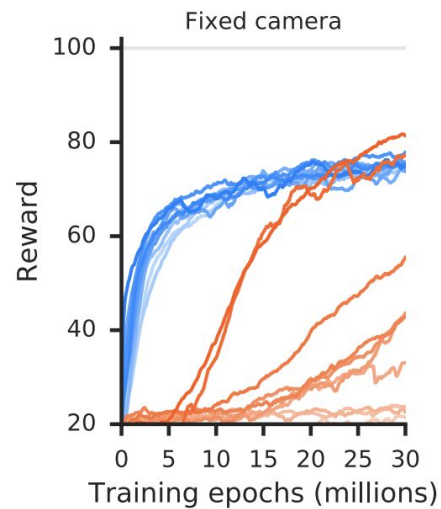
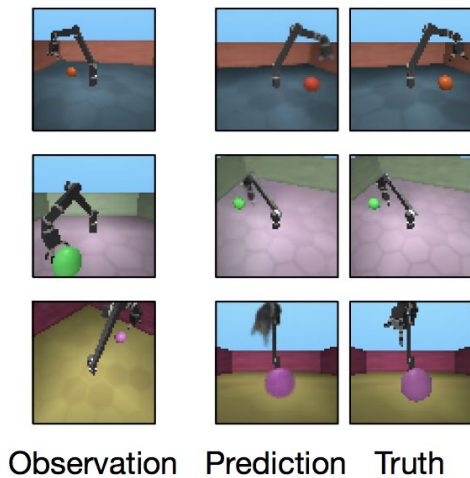
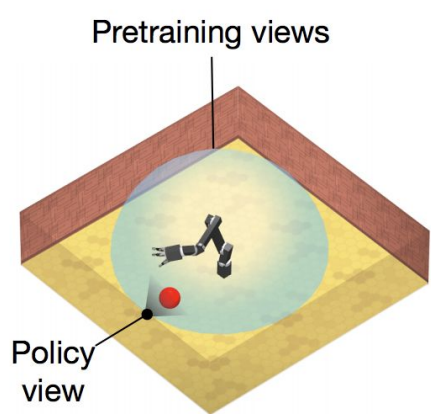
Observation



Samples



# GQN in reinforcement learning



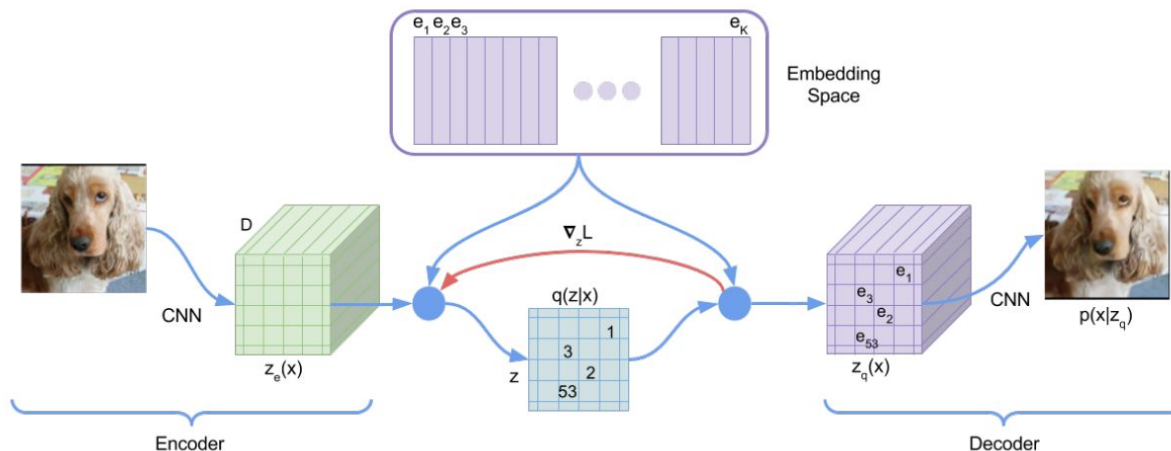
# Vector quantized VAEs (VQ-VAE)

Want to learn more?



Neural Discrete  
Representation Learning  
van den Oord et al., NIPS  
2017

Learning discrete latent variables is challenging (high variance gradient estimation).



Solution: reconstruct by using discrete latent variables to index into a learned continuous embedding space.



# VQ-VAE

Discrete latent variables can be used to capture high and low level information from the data.

Want to learn more?



Neural Discrete  
Representation Learning  
van den Oord et al, NIPS  
2017

*Data.*



*VQ-VAE reconstructions.*



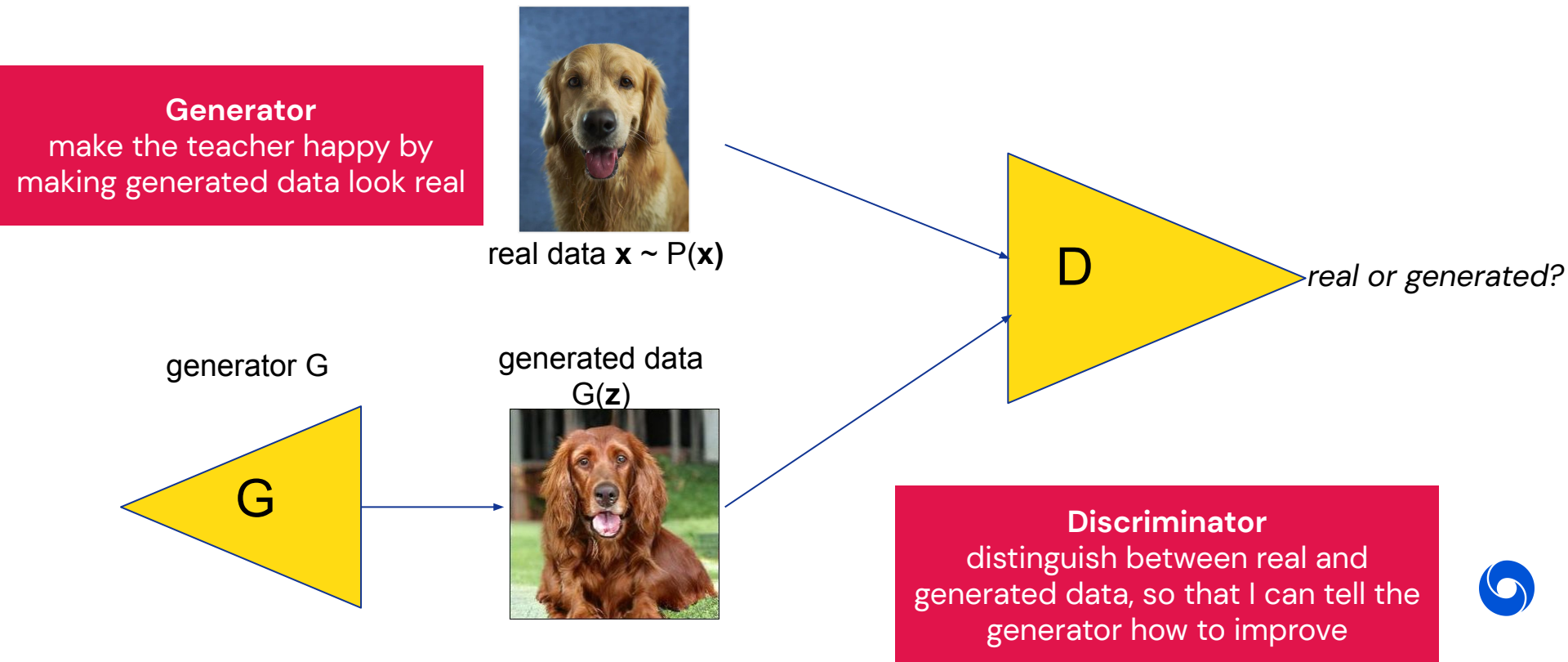
Figure from van den Oord et al. (2017)

# Generative adversarial networks

Want to learn more?



Goodfellow, et al. *Generative adversarial networks*.  
Neural Information  
Processing Systems (2014)



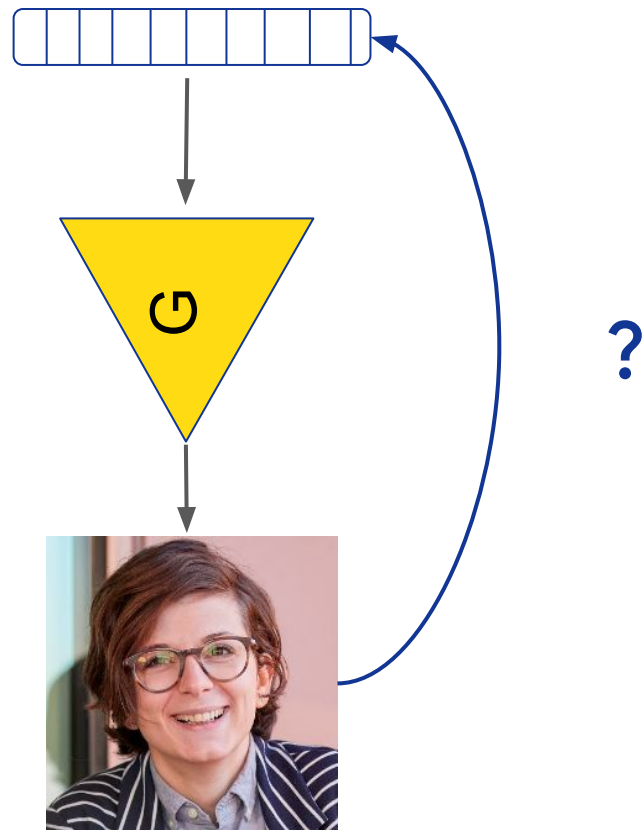
# Adversarially learned inference

Generation with GANs:

- No reconstruction loss.
- Learned model is implicit.

Inference with GANs:

- New model required.
- No uncertainty around learned representations.



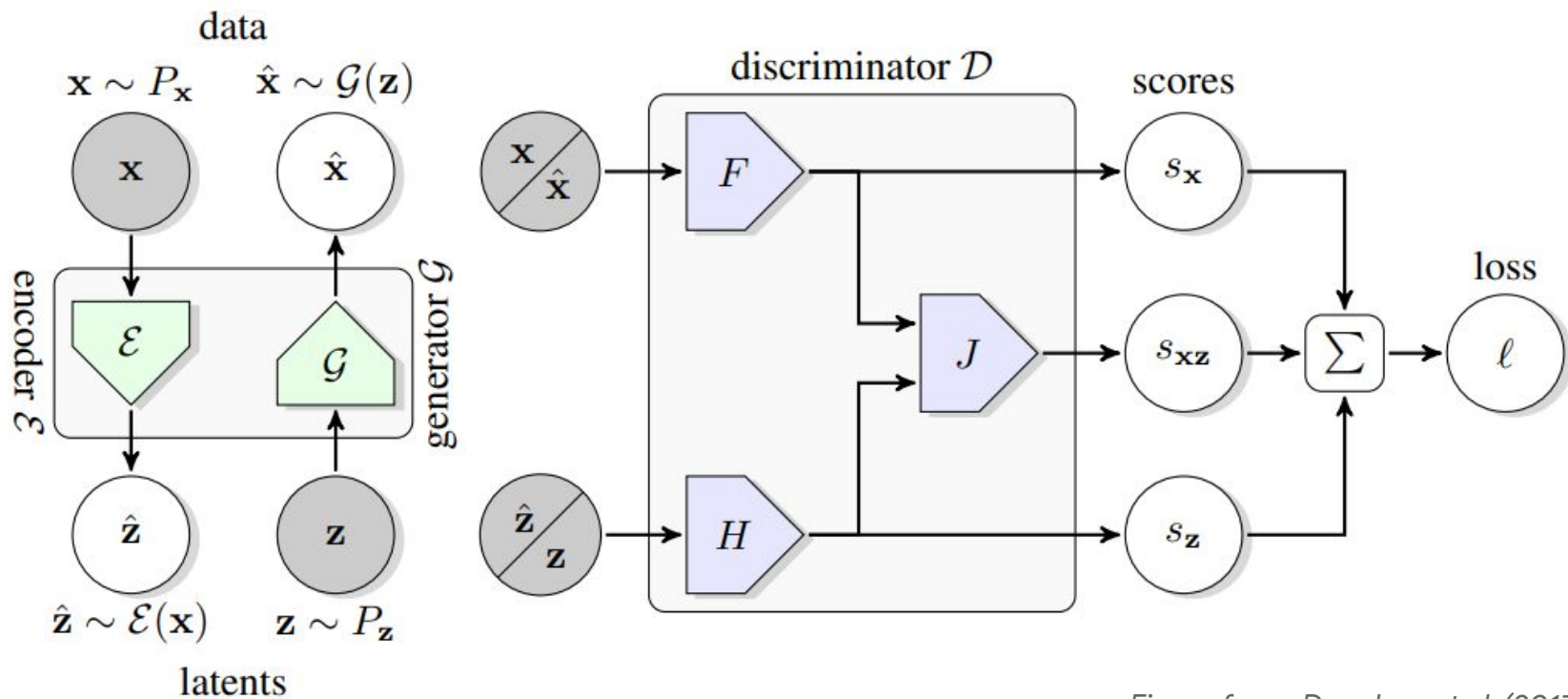


Figure from Donahue et al. (2017)

# BigBiGAN

- No pixel loss reconstruction → reconstructions capture high level information → latents capture high level information
- Latent representations extract meaningful features for semi supervised learning
  - Imagenet SOTA at time of publishing

Want to learn more?



Donahue, et al. Large Scale Adversarial Representation Learning. Neural Information Processing Systems (2019)



Figure from Donahue et al. (2017)



# GPT

Large scale generative models learn representations used for multiple downstream tasks.

Key: Neural architecture, billions of parameters and large amounts of data

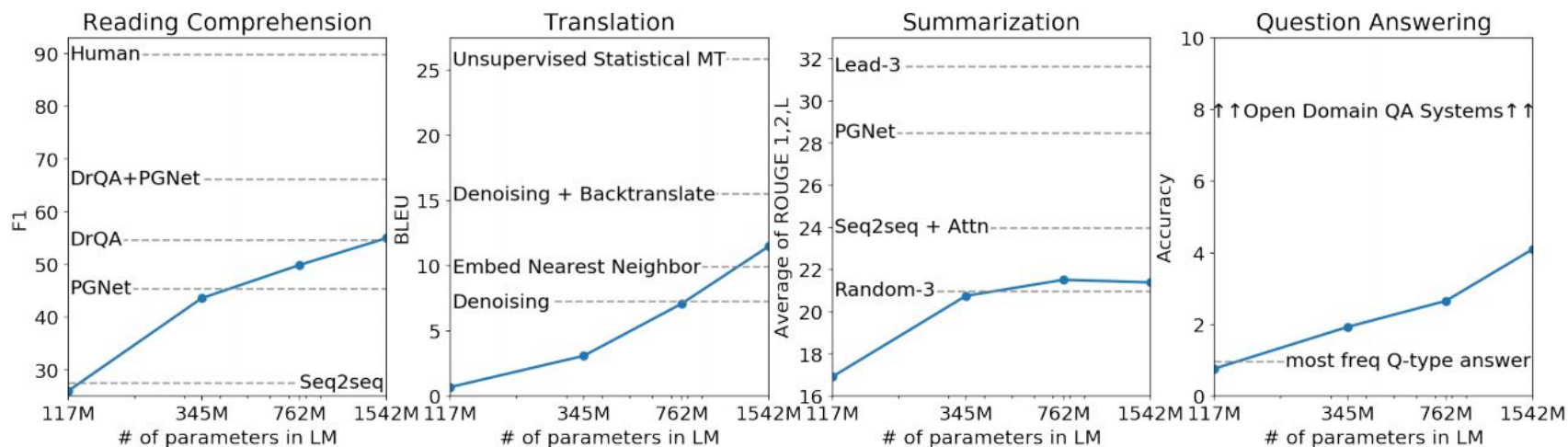
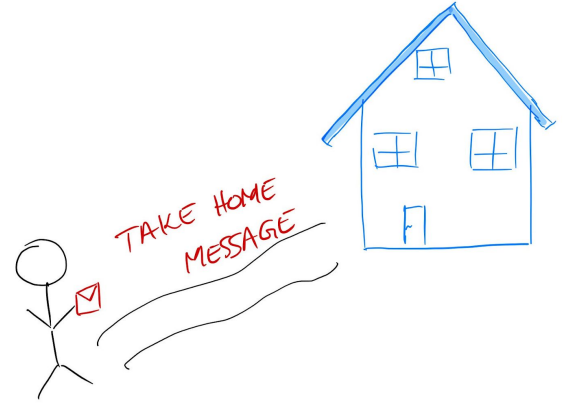


Figure from Radford et al. (2019)

**Latent variable models are a powerful tool for representation learning.**



# Contrastive learning





# Contrastive learning

- Removes the need for a generative model
- Ensure model learns the right context
- Use classification losses to learn representations



# Contrastive losses - word2vec

Want to learn more?



Mikolov, et al Distributed Representations of Words and Phrases and their Compositionality. NIPS (2013)

- Predict which words appear next.
- Contrastive: provide positive examples of future words.
- Contrastive: provide negative examples of words which won't come next.

$$\log \sigma(v'_{w_O} \top v_{w_I}) + \sum_{i=1}^k \mathbb{E}_{w_i \sim P_n(w)} \left[ \log \sigma(-v'_{w_i} \top v_{w_I}) \right]$$

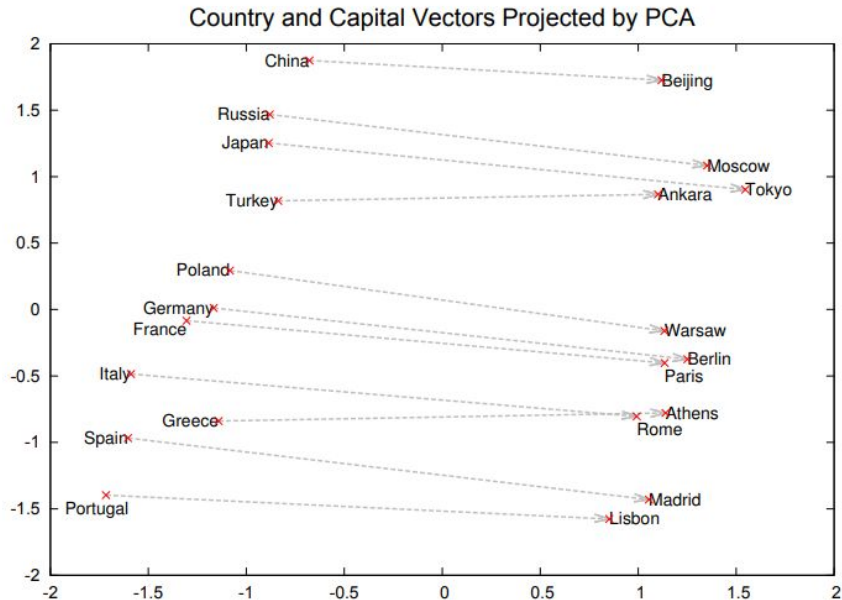


Figure from Mikolov et al. (2013)

# Few shot machine translation with word2vec

Learn a dictionary:

- Use word2vec to learn word representations
- Use a few hundred examples to learn a linear mapping between words
- Can now translate!

Want to learn more?



Mikolov, et al Exploiting Similarities among Languages for Machine Translation arxiv(2013)

English word	Computed Spanish Translation	Dictionary Entry
pets	mascotas	mascotas
mines	minas	minas
unacceptable	inaceptable	inaceptable
prayers	oraciones	rezo
shortstop	shortstop	campocorto
interaction	interacción	interacción
ultra	ultra	muy
beneficial	beneficioso	beneficioso
beds	camas	camas
connectivity	conectividad	conectividad
transform	transformar	transformar
motivation	motivación	motivación



Table from Mikolov et al. (2013)

# Contrastive predictive coding

- maximize mutual information between data and learned representations
- uses supervised learning to model density ratios

$$\mathcal{L}_N = -\mathbb{E}_X \left[ \log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)} \right]$$

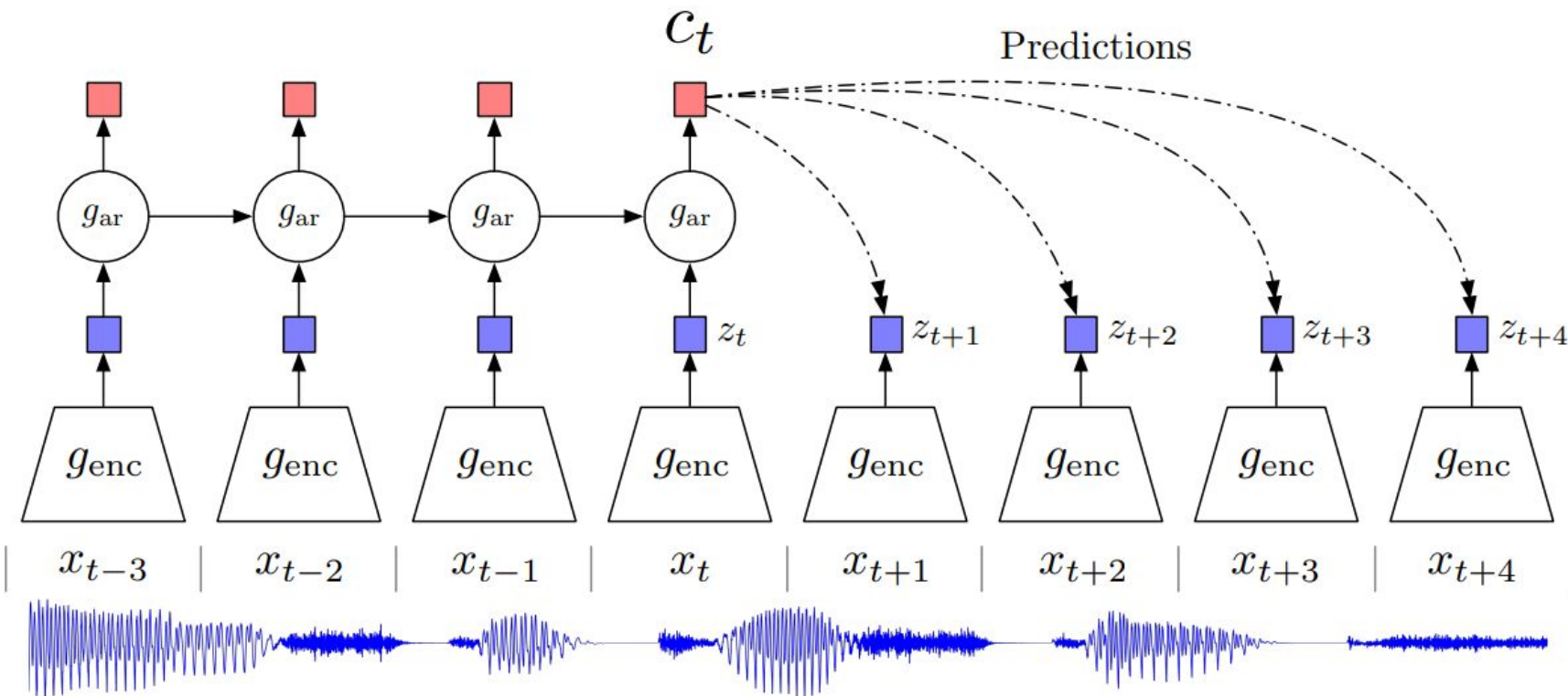
Want to learn more?



van den Oord, et al  
Representation Learning with  
Contrastive Predictive  
Coding arxiv(2018)



# Contrastive predictive coding



# Contrastive predictive coding

Learning from contrastive representations (learned unsupervised) is more efficient in the low data regime compared to learning from pixels.

Want to learn more?



Hénaff, et al Data-efficient image recognition with contrastive predictive coding arxiv(2019)

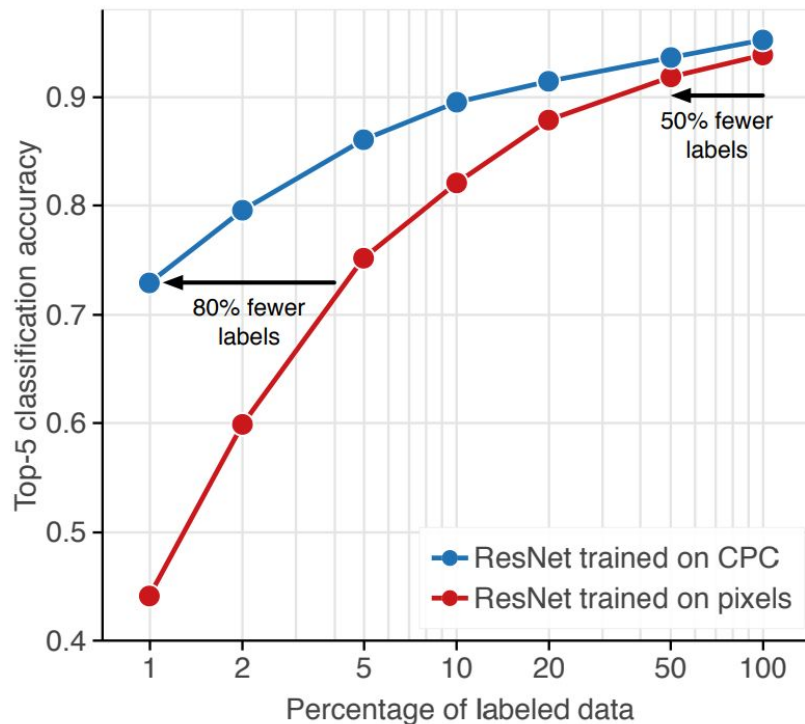


Figure from Hénaff et al. (2019)

# SimCLR

Use contrastive losses to maximize mutual information between representations of data under different transformations.

Want to learn more?



Chen, et al. A Simple Framework for Contrastive Learning of Visual Representations arxiv(2020)

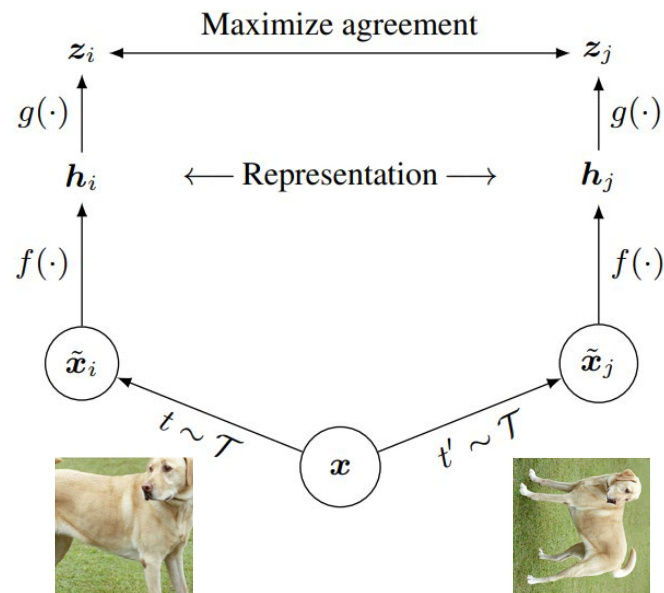


Figure from Chen et al. (2020)



(a) Original



(b) Crop and resize



(c) Crop, resize (and flip)



(d) Color distort. (drop)



(e) Color distort. (jitter)



(f) Rotate  $\{90^\circ, 180^\circ, 270^\circ\}$



(g) Cutout



(h) Gaussian noise



(i) Gaussian blur



(j) Sobel filtering





# SimCLR

Achieves SOTA on Imagenet benchmarks, both on a linear classifier trained on input representations as well as semi supervised learning (10% improvement over previous state of the art).

Want to learn more?



Chen, et al. A Simple Framework for Contrastive Learning of Visual Representations arxiv(2020)

Linear classifier with input given by learned representations.

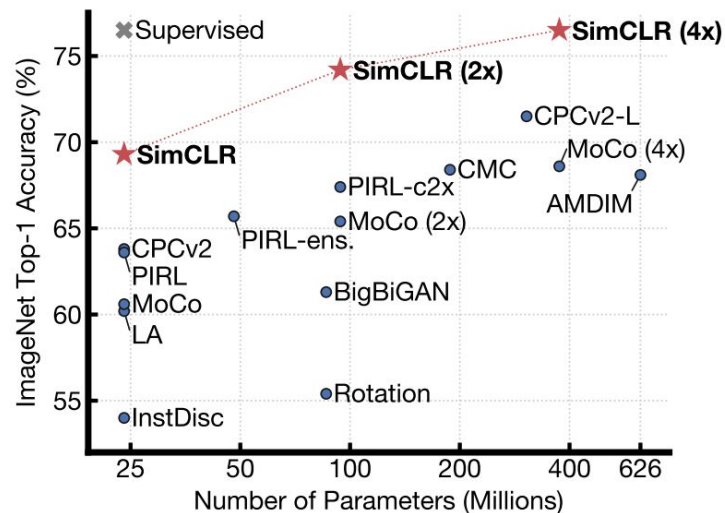
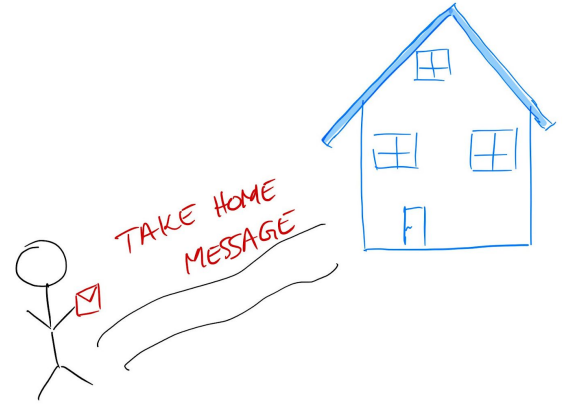


Figure from Chen et al. (2020)

**Contrastive losses use classifiers to learn representations which are temporally or spatially consistent.**

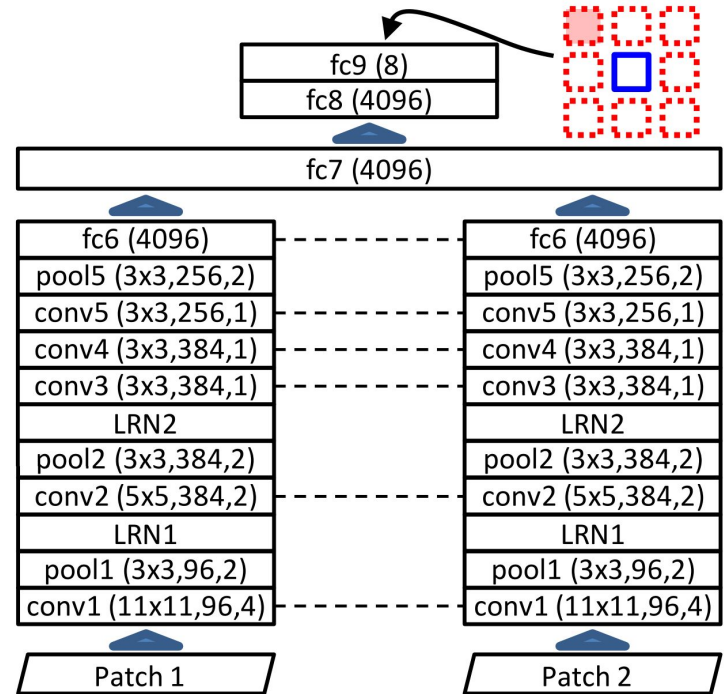


# Self supervised learning



# Self supervised learning

- Design tasks based on data modalities
- Easy to obtain data
- Representations are deterministic, often features of neural networks



# Colorful Image Colorization

Want to learn more?



Zhang, et al Colorful Image  
Colorization arxiv (2016)

- Task: colourize image.
- Easy to obtain supervised data.
- Representations useful for semi supervised learning (Imagenet).



Figure from Zhang et al. (2016)

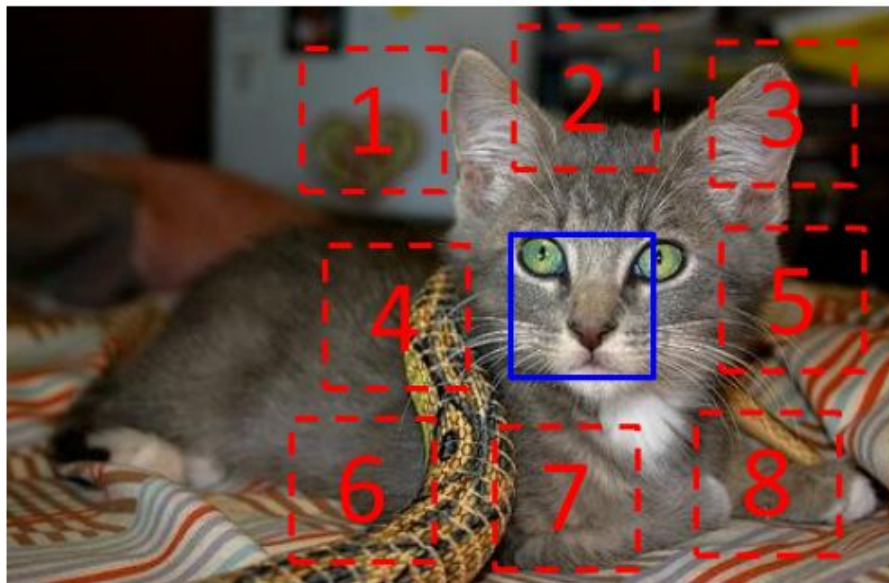
# Unsupervised Visual Representation Learning by Context Prediction

Want to learn more?



Doersch, et al Unsupervised  
Visual Representation  
Learning by Context  
Prediction ICCV (2015)

- Task: predict selected patches.
- Features useful in semi supervised learning.
- Unsupervised object discovery.



$$X = \left( \begin{array}{c|c} \text{[Face Patch]} & \text{[Ear Patch]} \end{array} \right); Y = 3$$

Figure from Doersch et al. (2015)

# Unsupervised Representation Learning by Sorting Sequences

Want to learn more?



Lee et al Unsupervised Representation Learning by Sorting Sequences IEEE (2017)

- Learn temporal coherence from video.
- Avoids data generation in pixel space.
- Used for semi supervised learning – object recognition, action recognition.

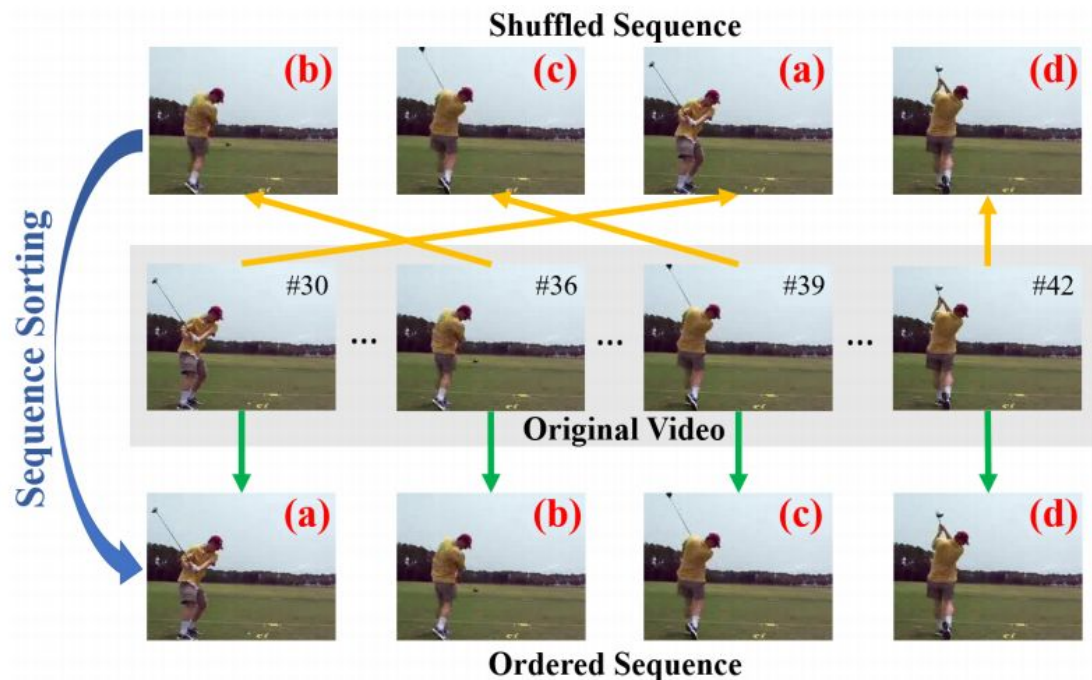


Figure from Lee et al (2017)

# BERT

Want to learn more?



Devlin, et al BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding ACL(2019)

- Representations learned with multiple tasks (self supervision meets generative modelling)
- Predict missing tokens (past and present)
- Classify which sentence order
- Sparked a revolution in NLP

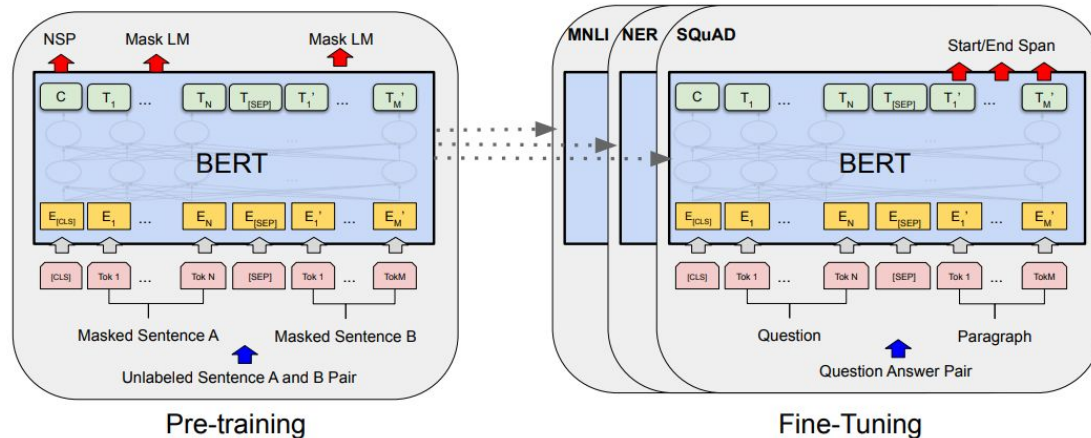


Figure from Devlin et al. (2019)



# BERT

Want to learn more?



Devlin, et al BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding ACL(2019)

Used for multiple downstream tasks:

- Summarization.
- Named entity recognition.
- Spam detection.
- In production: Part of Google search.

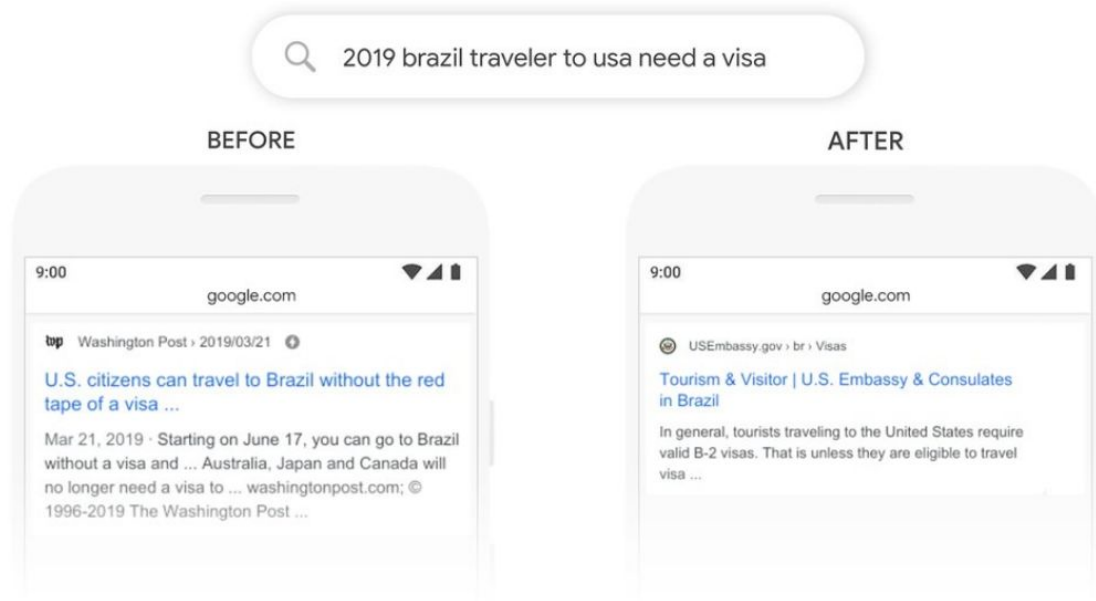
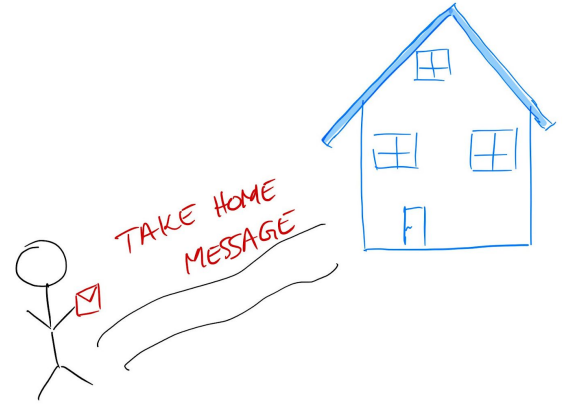


Figure from Google blogpost by Nayak et al. (2019)

**Self supervised learning exploits domain knowledge to build tasks useful for representation learning.**



## Keep in mind that....

- Task design for learning representations is important
- Modality is important
- Context is important
- Learning generative models is hard, might be able to get away without it (contrastive losses, self supervision)
- Crucial benefits by incorporating changes in neural architectures



DeepMind

6

Future



## Next...

- Generative models: Powerful posteriors and better priors.
- Contrastive learning: going beyond temporal and spatial coherence.
- Self supervised learning: more task design.
- Incorporating changes in neural representations.
- Causality.



**Thank you**





# Questions



# TC-VAE







# Upcoming lectures

- Attention and Memory in Deep Learning
- Generative Latent Variable Models and Variational Inference
- Responsible innovation

