DeepMind

Representation Learning Without Labels

Irina Higgins @irinavlh
Danilo J. Rezende @danilojrezende
S. M. Ali Eslami @arkitus

ICML 2020

DeepMind

Acknowledgements

Mihaela Rosca, Shakir Mohamed, Alex Graves,
Olivier Henaff, Brian McWilliams, Steven McDonagh,
David Pfau, Jovana Mitrovic, Andrew Zisserman

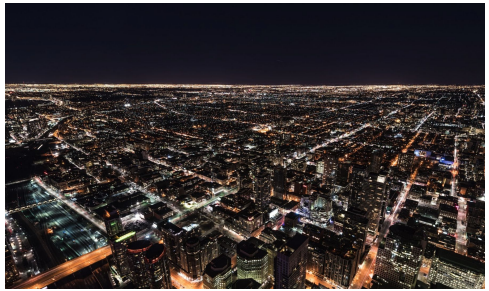
ICML 2020

Agenda for this tutorial

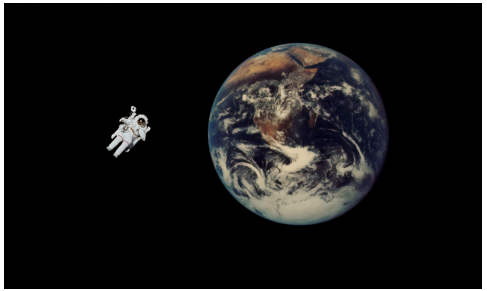
01
Introduction

02
Building Blocks

03
Landscape



04
Frontiers



Scope for this tutorial

What this tutorial is

- An overview of building blocks
- A comparison of methods
- Focus on the image modality

What this tutorial isn't

- A comprehensive list of all relevant techniques
- Representation learning for different modalities (text, audio, video, graphs, etc.)

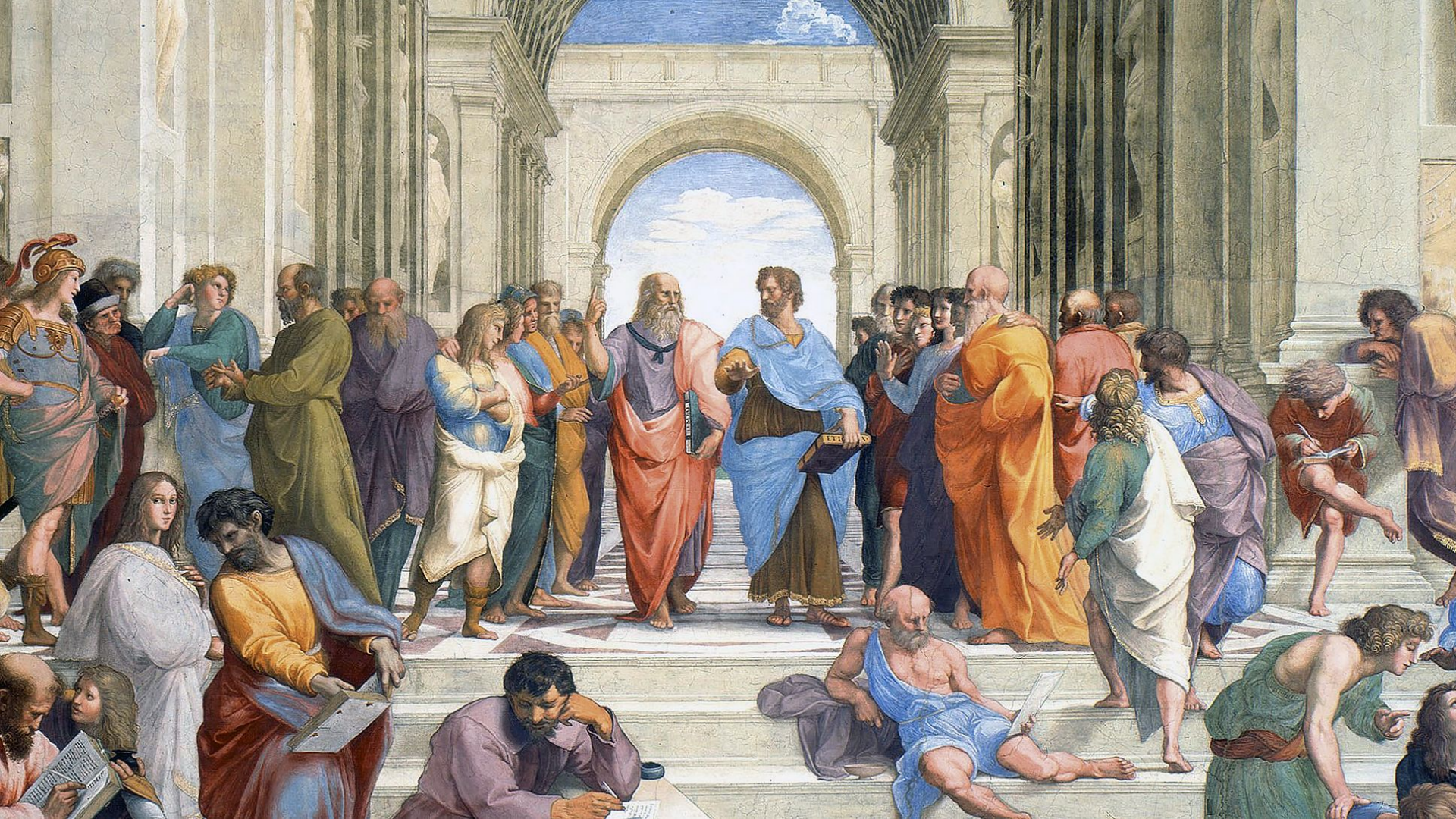
Disclaimers

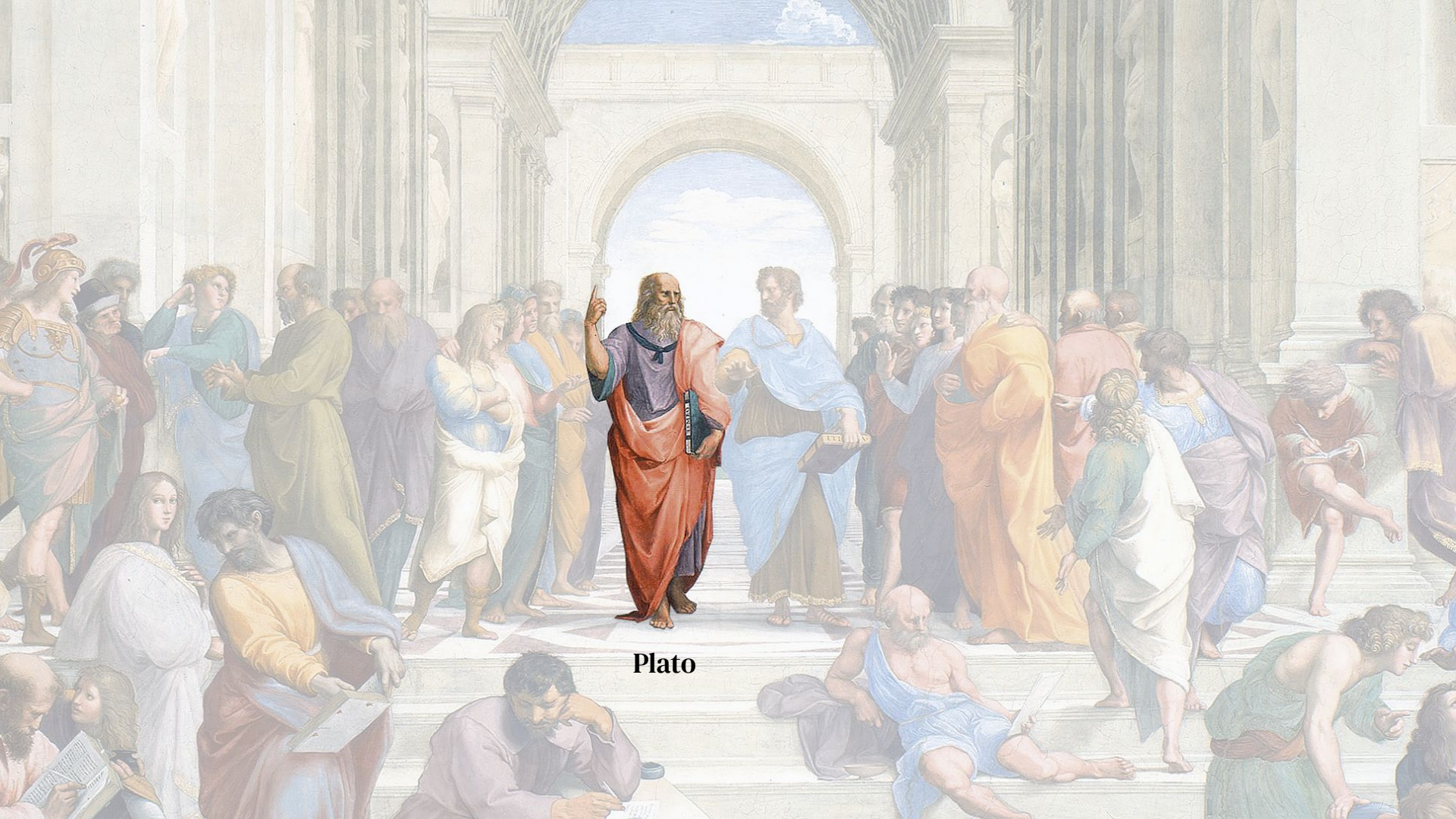
Want to learn more?

- This is a huge topic with a vast, multi-disciplinary history
- We will inevitably miss important related work
- Each citation is only meant as a representative example; seed for **connectedpapers.com**
- There are many views on the literature, this is one
- Not necessarily chronological
- Email us with pointers or suggestions and we will update the slides

DeepMind

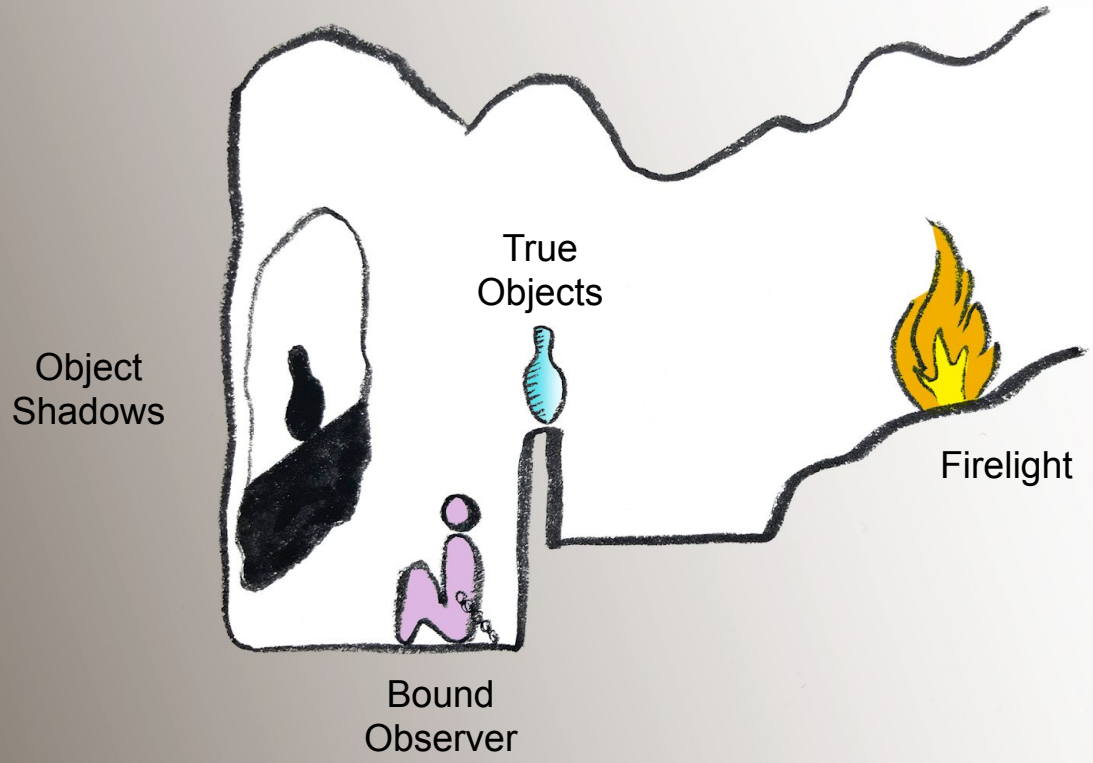
1 Introduction



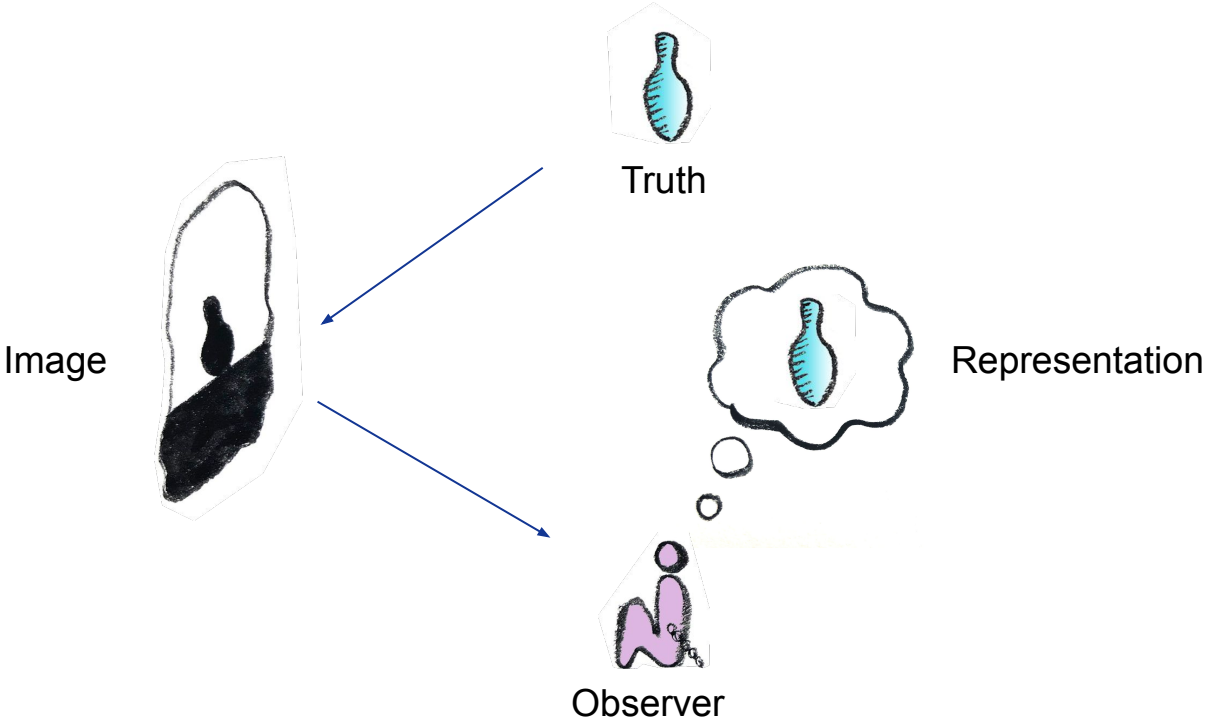


Plato

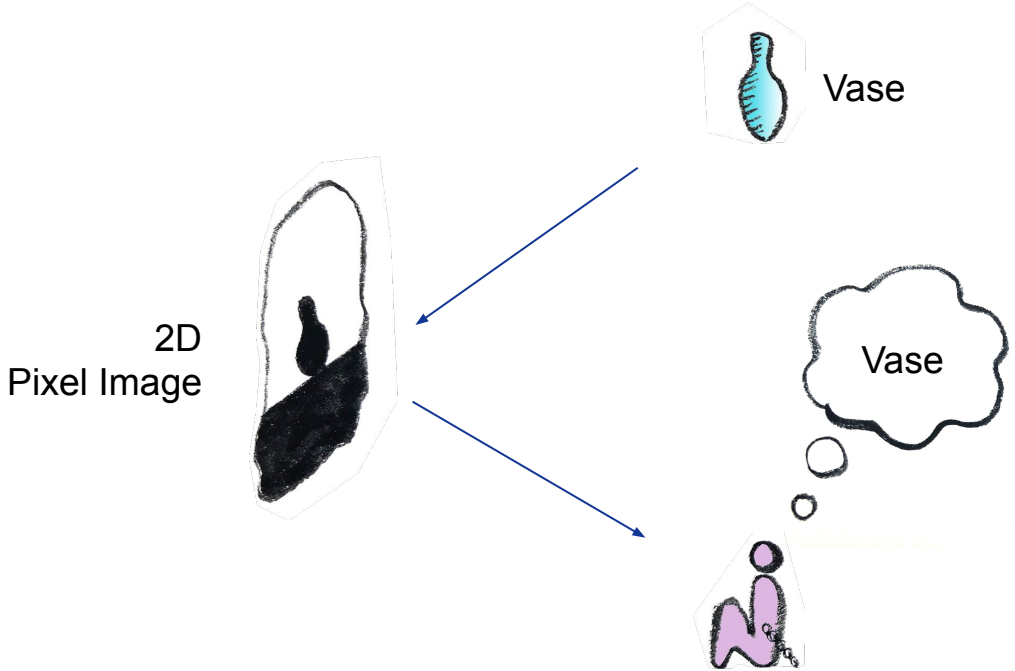
Plato's allegory of the cave



Plato's allegory of the cave

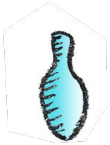


Desired understanding: simplistic view



The representation problem

2D
Pixel Image



- **Class:** Vase
- **Shape:** Gourd
- **Colour:** Blue
- **Height:** 15cm tall
- **Weight:** 230g
- **Scratched:** Yes
- **Moving:** No
- And many more attributes...

- Which attributes?
- What formats?
- Partial observability?
- How quickly?
- Measure of success?

“

Vision is a process that produces from images of the external world a description that is useful to the viewer and not cluttered by irrelevant information.

— Marr and Nishihara, 1978

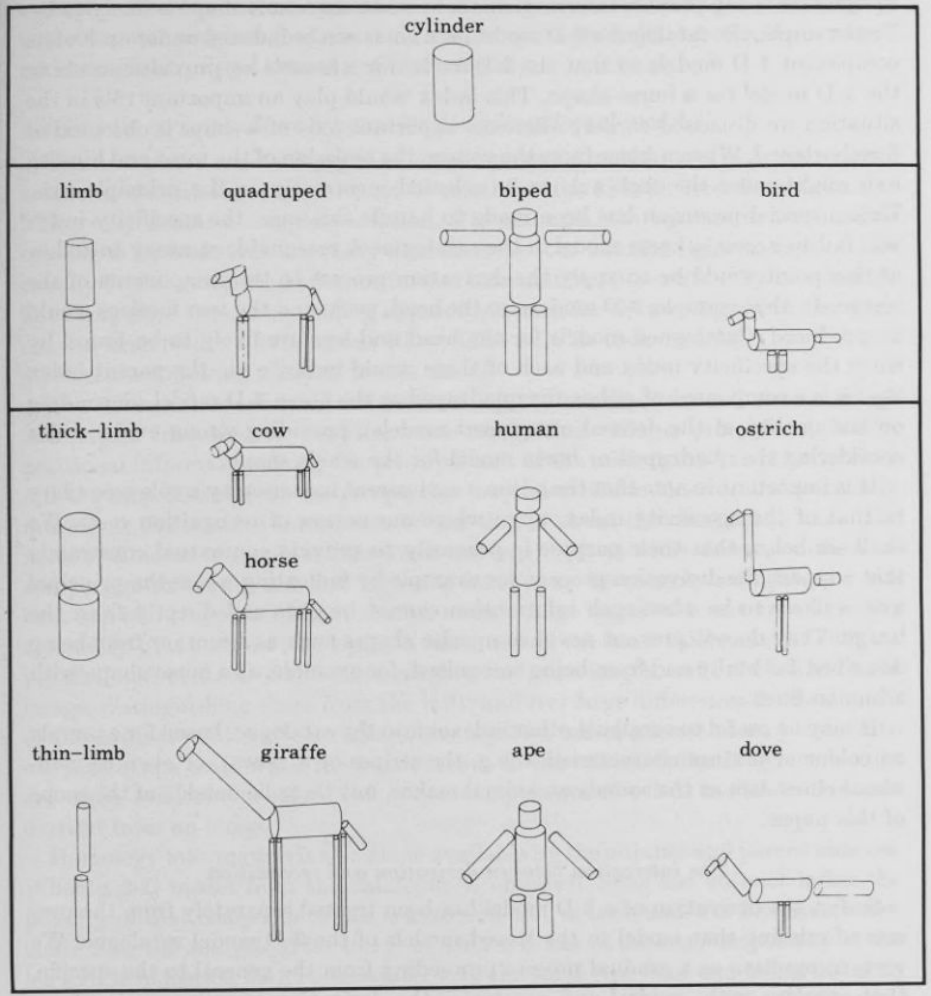
Want to learn more?

Representation and Recognition of the Spatial Organization of Three-Dimensional Shapes, Marr et al, Proc. R. Soc. Lond. (1978)

“

Representation is a 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?

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

- Representational **form** is orthogonal to information **content**
- Useful **abstraction** to make different computations and tasks **more efficient**

The supervised solution

1. Decide **which attributes** you care about
2. Decide the **format** for each attribute
3. Create a **large dataset** of (image, label) pairs
4. **Train a neural network** to predict labels

label= ‘Vase’ label= ‘Ball’ label= ‘Dog’

The supervised solution

1. Decide **which attributes** you care about
2. Decide the **format** for each attribute
3. Create a **large dataset** of (image, label) pairs
4. **Train a neural network** to predict labels

label= ‘Vase’ label= ‘Ball’ label= ‘Dog’

Works extremely well on a diverse set of problems

However, it also raises several questions:

1. Who provides ground truth for the labels?
2. What if we don’t ‘know’ the groundtruth ourselves?
3. Which attributes are chosen for labelling?
4. Which attributes are ignored for labelling?
5. What biases do the labels propagate?
6. Do children learn purely from labels?
7. Do animals learn from labels at all?
8. **Can useful representations develop without labels?**

Can useful representations develop without labels?

If the answer is yes:

1. We learn more efficiently when we do gain access to labels

2. We still learn useful things when label collection is impossible

Supervised vs Unsupervised
Labelled vs Unlabelled

Often it is difficult to formally distinguish between **supervised** and **unsupervised** techniques

For example:

- Image captioning (specification of **caption**)
- Machine translation (specification of **pairing**)
- Reinforcement learning (specification of **reward**)
- Generative models (specification of **structure**)
- Language modeling (specification of **dataset**)

Objective is to reduce reliance on labels **that can only be assigned by human brains**, and learn more from **raw measurements of the world**

Still, the term ‘unsupervised’ is actually rather convenient

Universal Dictionary of Arts, Sciences and Literature, 1810

History of representation learning

Want to learn more?

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

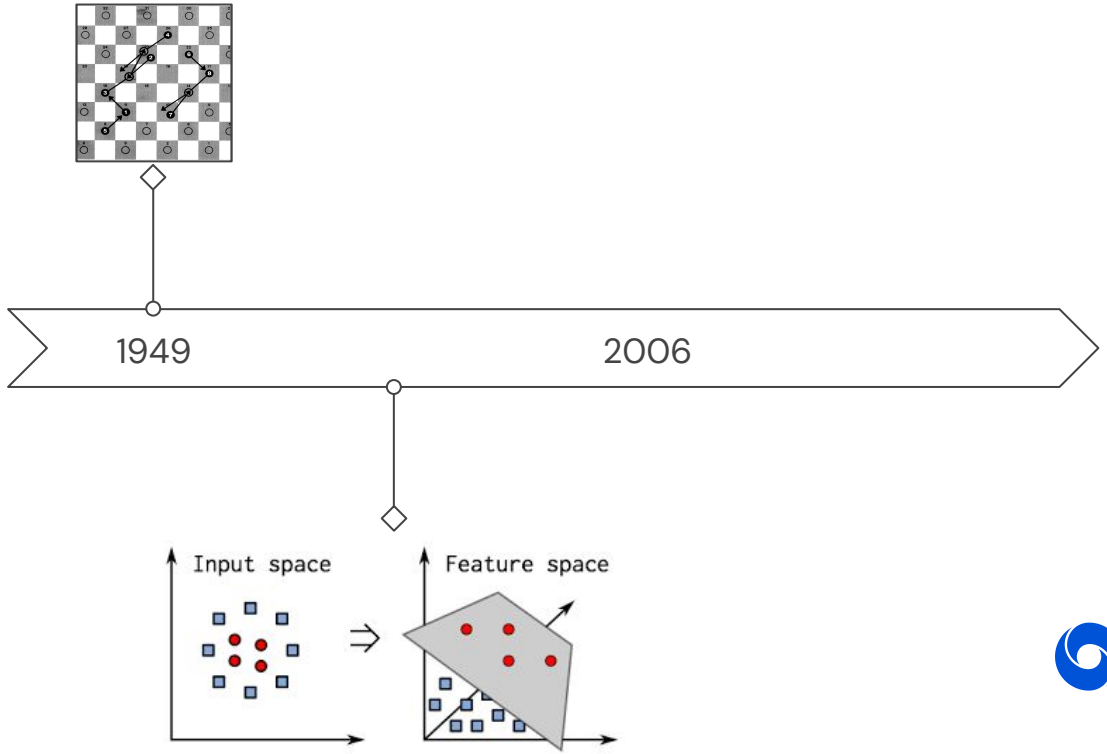
- Arthur Samuel coins the term “machine learning”

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

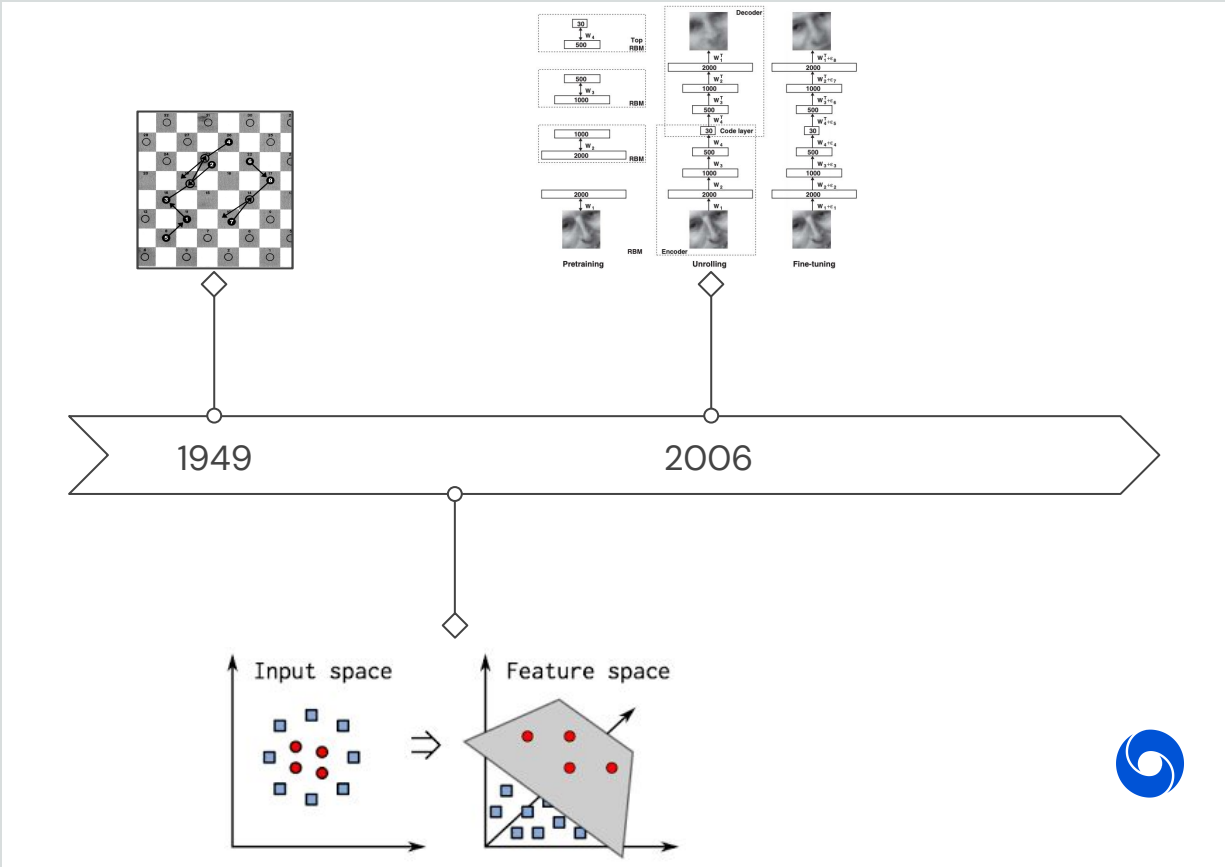


History of representation learning

Want to learn more?

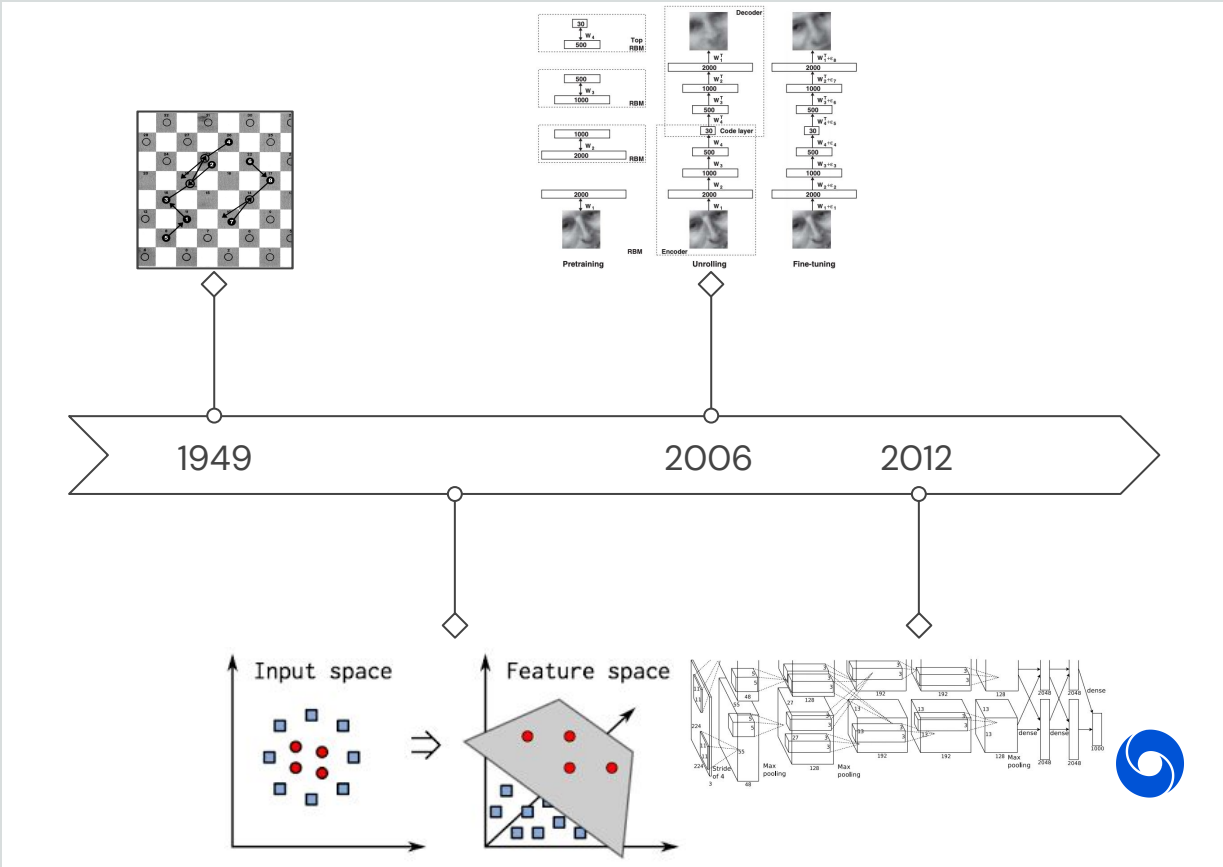
Reducing the Dimensionality of Data with Neural Networks, Hinton and Salakhutdinov, Science (2006)

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



History of representation learning

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



Want to learn more?

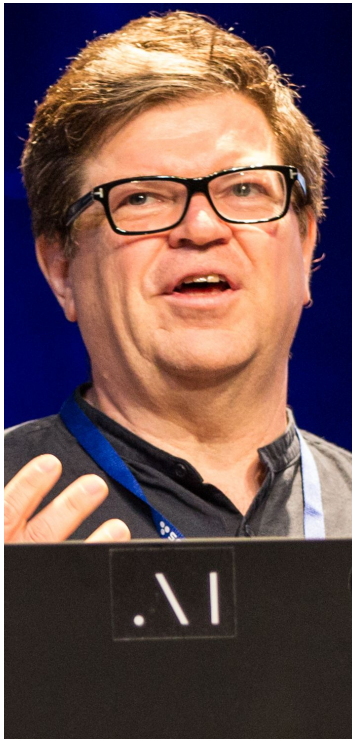
ImageNet Classification with Deep Convolutional Neural Networks, Krizhevsky et al, NeurIPS (2012)

Turing Award winners at AAAI 2020

> I always knew unsupervised learning was the right thing to do
>
> — Geoff Hinton

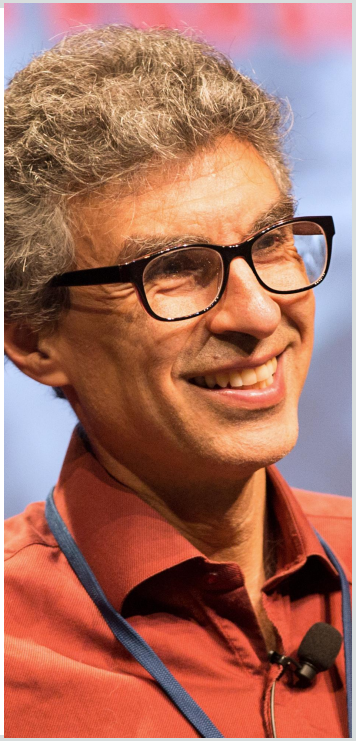
> Basically it's the idea of learning to represent the world before learning a task — and this is what babies do
>
> — Yann LeCun

> 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



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

Eviatar Bach / CC BY-SA



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

Intelligence without representation

> There is no clean division between perception (abstraction) and reasoning in the real world. The brittleness of current AI systems attests to this fact.
>
> — Rodney Brooks

Science History Institute / CC BY-SA

Want to learn more?

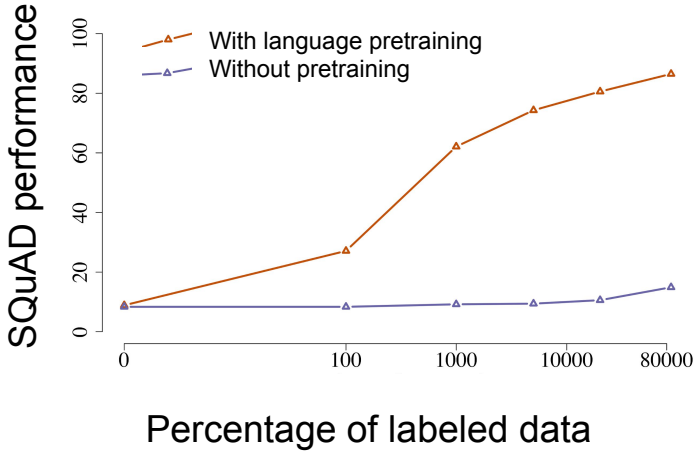
Intelligence without representation, Rodney A. Brooks, Artificial Intelligence (1991)

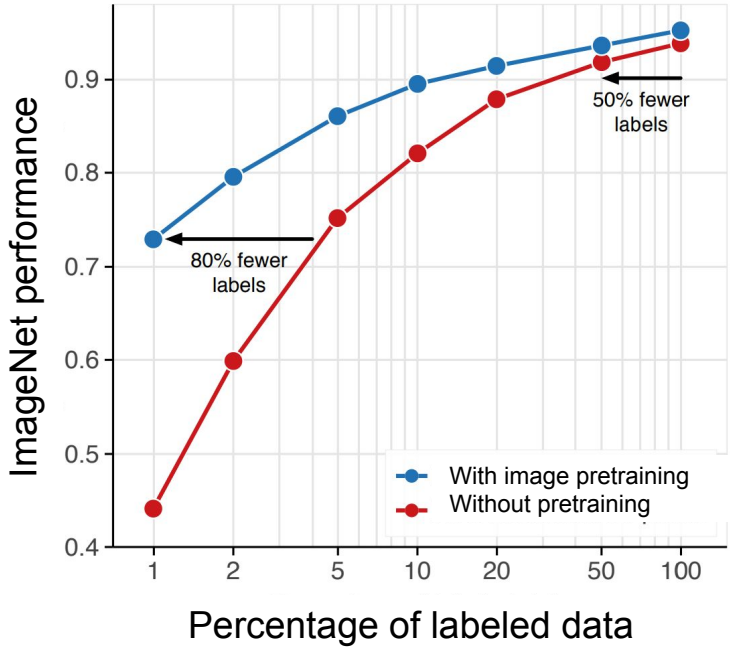
Impressive recent progress

Want to learn more?

Learning and Evaluating General Linguistic Intelligence, Yogatama et al (2019)

Data-Efficient Image Recognition with Contrastive Predictive Coding, Olivier J. Hénaff et al, ICML (2020)





DeepMind

2 Building Blocks

The representation problem

Realm of **Images**

Representation
(Understanding)

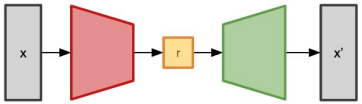
Generation
(Imagination)

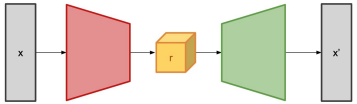
Realm of **Representations**

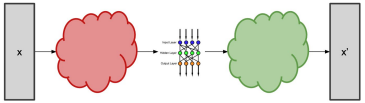
Model zoo

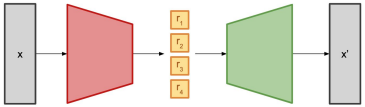
Curtis's Botanical Magazine, 1831

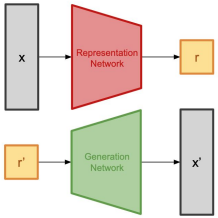
Model zoo

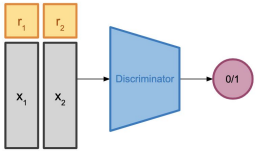

x
r
x'

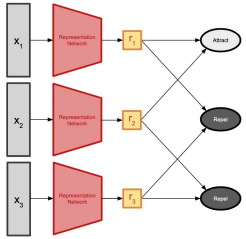

x
r
x'

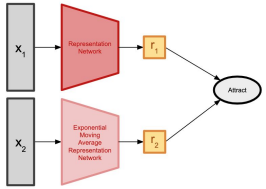

x
x'

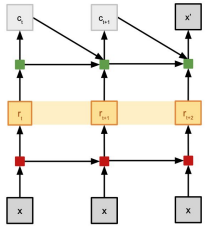

x
x'

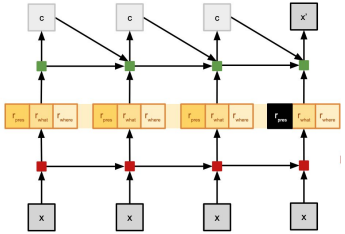

x
Representation Network
r
r'
Generation Network
x'

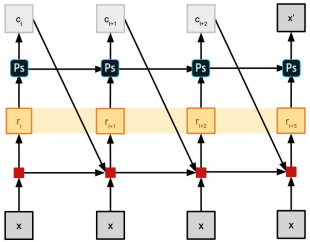

r1
r2
x1
x2
Discriminator
0/1


x1
x2
x3
Representation Network
r1
r2
r3
Attract
Repel
Repel

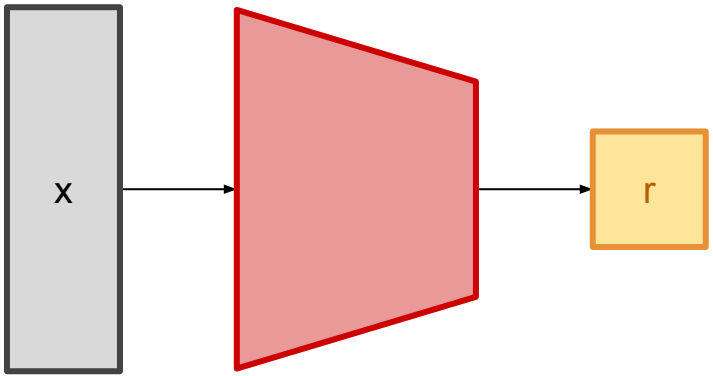

x1
x2
Representation Network
Exponential Moving Average Representation Network
r1
r2
Attract


x'
x
x
x


c
c
c
x'
x
x
x
x


Ps
Ps
Ps
Ps
x'
x
x
x
x

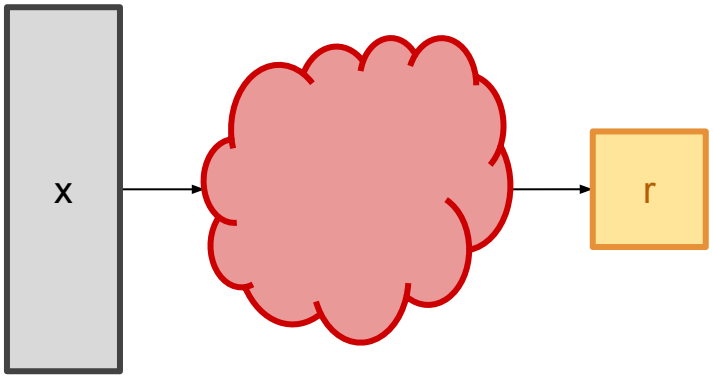
(Representation / Encoder / Inference) Networks



- **Size:** Smaller or larger than x
- **Structure:** Flat or interpretable
- **Type:** Continuous or discrete
- **Shape:** Fixed or variable
- **Disentangled** or not

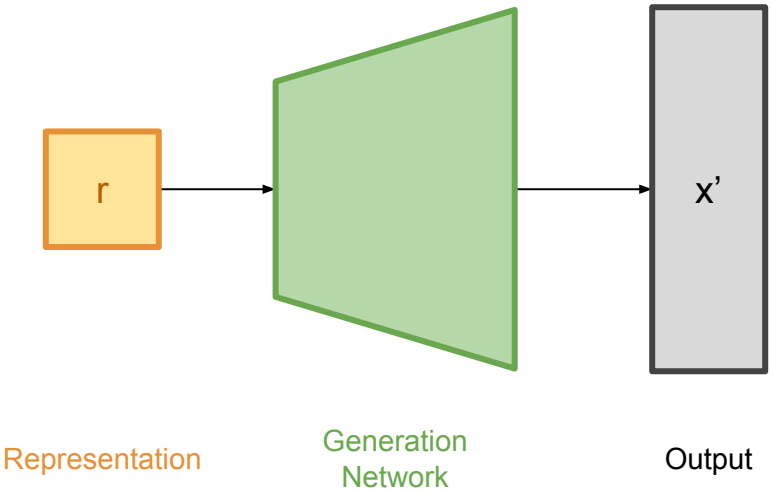
- Multi-layer perceptron
- ConvNet
- Transformer
- Recurrent neural net

(Representation / Encoder / Inference) Networks



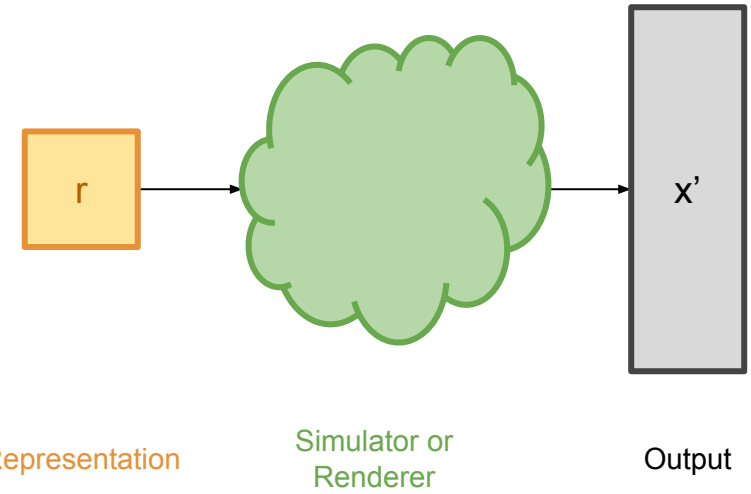
- Differentiable or not
- Interpretable or not
- Deterministic or stochastic

(Generation / Generator / Decoder) Networks



- Multi-layer perceptron
- DeconvNet
- Transformer
- Recurrent neural net

(Generation / Generator / Decoder) Networks



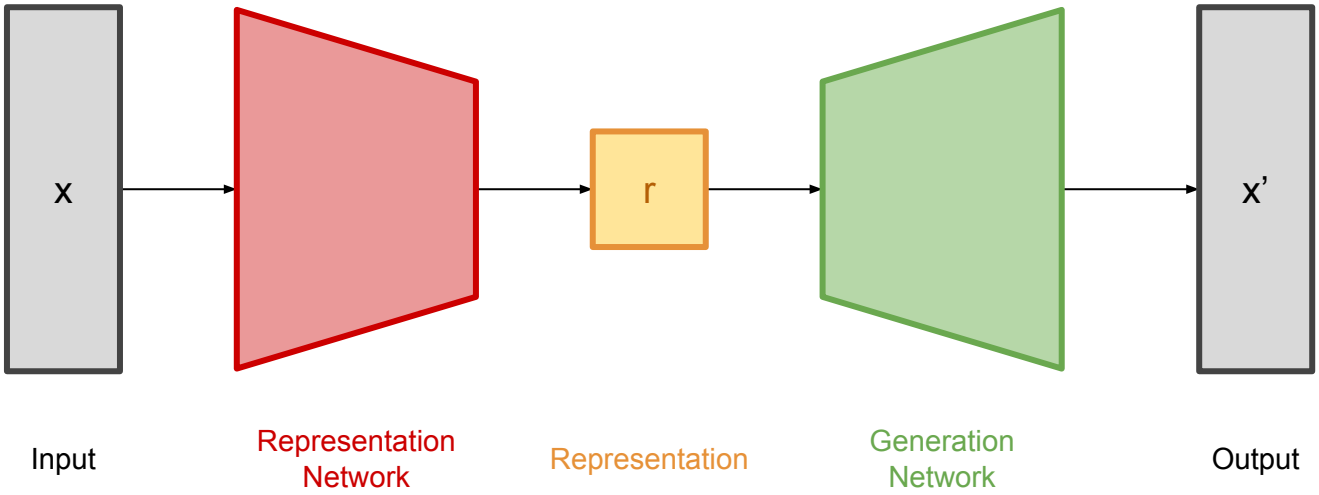
- Differentiable or not
- Interpretable or not
- Deterministic or stochastic

Autoencoders

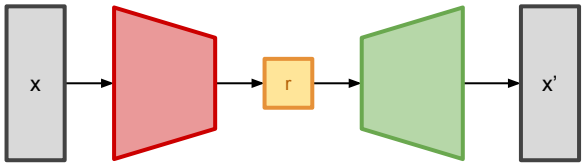
Want to learn more?

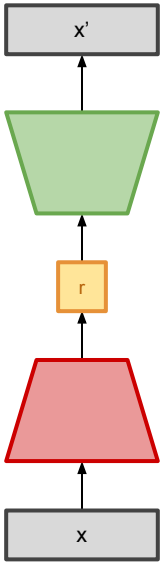
Auto-Encoding Variational Bayes, Kingma et al, ICLR (2014)

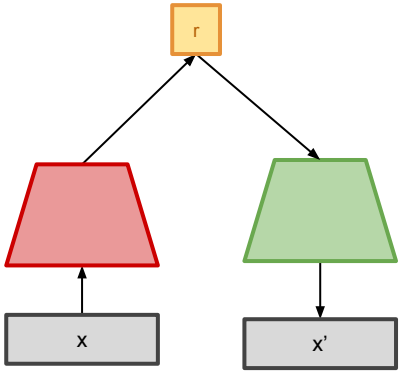
Stochastic backpropagation and approximate inference in deep generative models, Rezende et al, ICML (2014)



Autoencoder Graphics





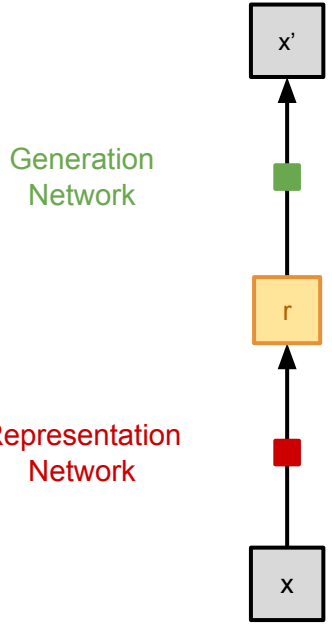


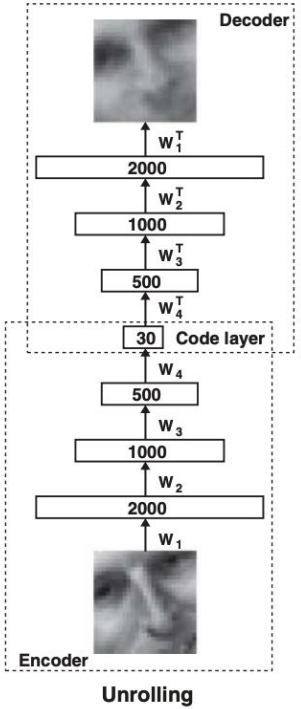
Autoencoders: What are they for?

- Density estimation
- Dimensionality reduction
- Image generation
- Denoising
- **Representation learning**

Want to learn more?

Reducing the Dimensionality of Data with Neural Networks, Hinton et al, Science (2006)



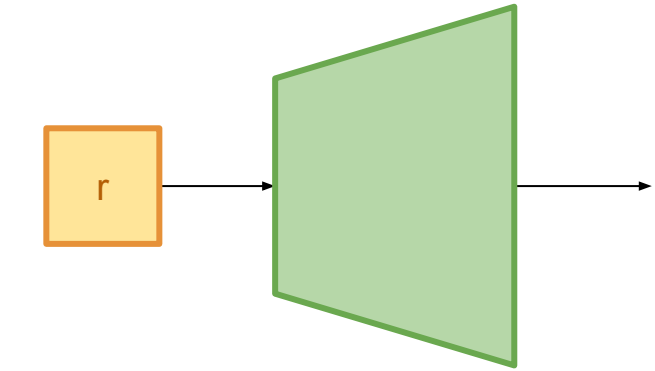


Unrolling

Autoencoders

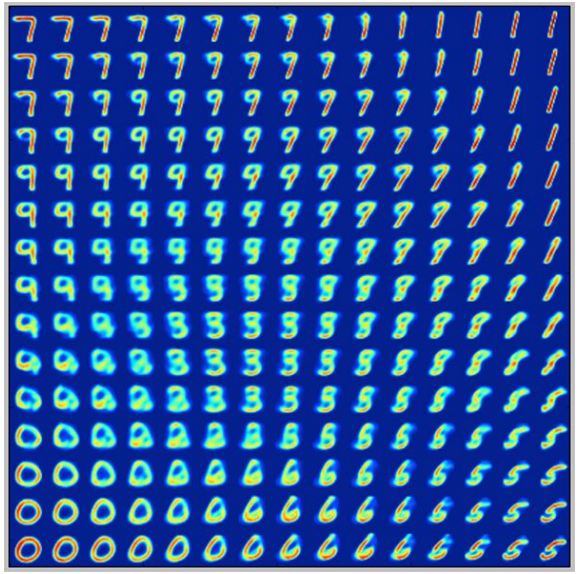
Want to learn more?

Building Autoencoders in Keras, Chollet (2016)



Representation

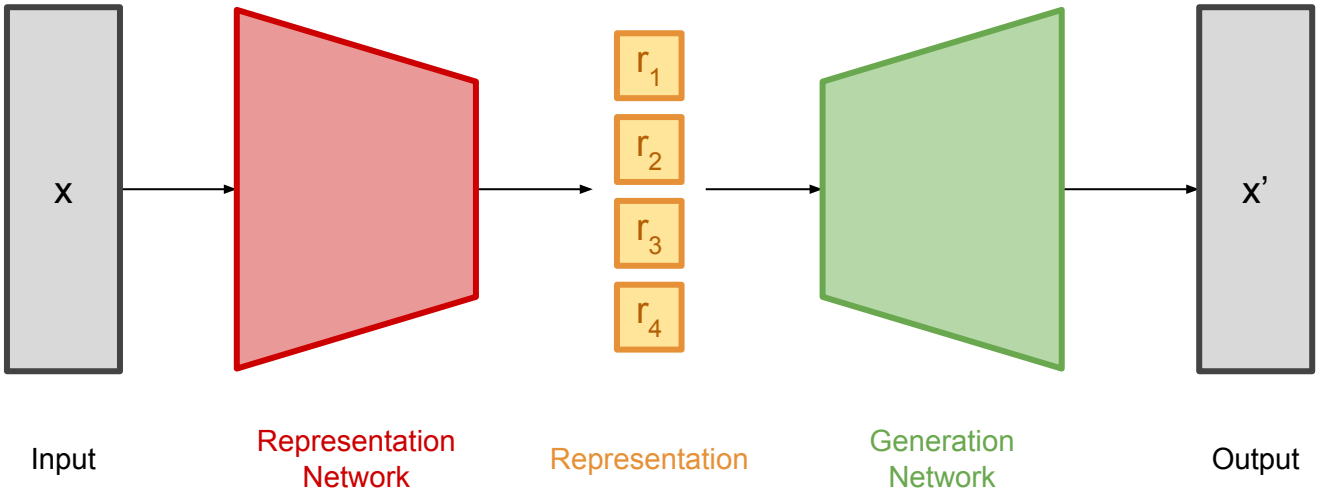
Generation Network



Disentangled Autoencoders

Want to learn more?

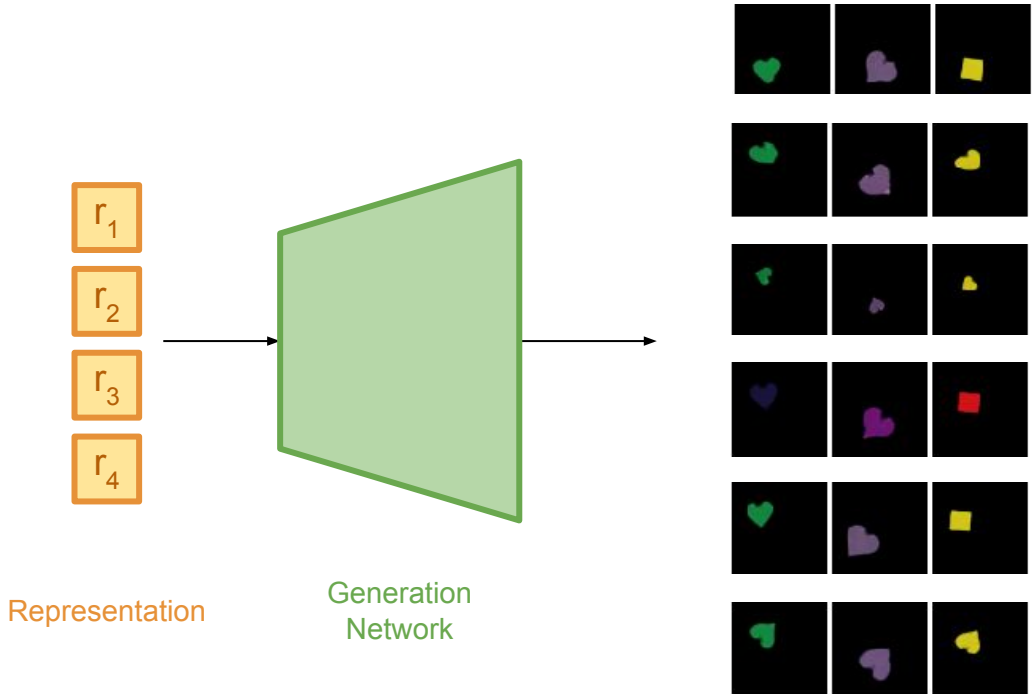
β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework, Higgins et al., ICLR 2017



Disentangled Autoencoders

Want to learn more?

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



GIFs adapted from Chris Burgess

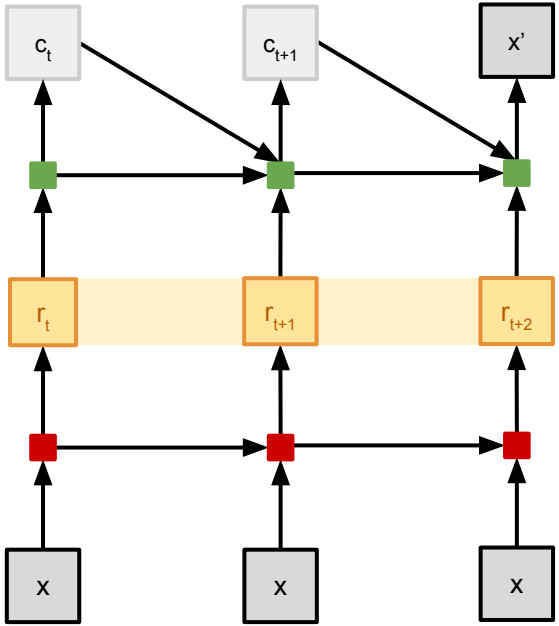
Sequential Autoencoders

Want to learn more?

Towards Conceptual Compression, Gregor et al, NeurIPS (2016)

Generation Network

Representation Network



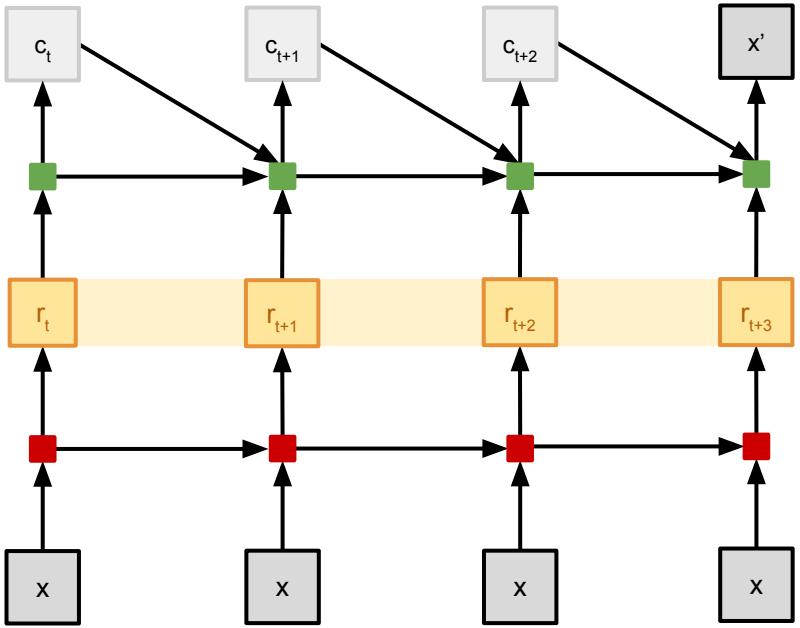
Generation RNN

Representation RNN

Sequential Autoencoders

Want to learn more?

Towards Conceptual Compression, Gregor et al, NeurIPS (2016)



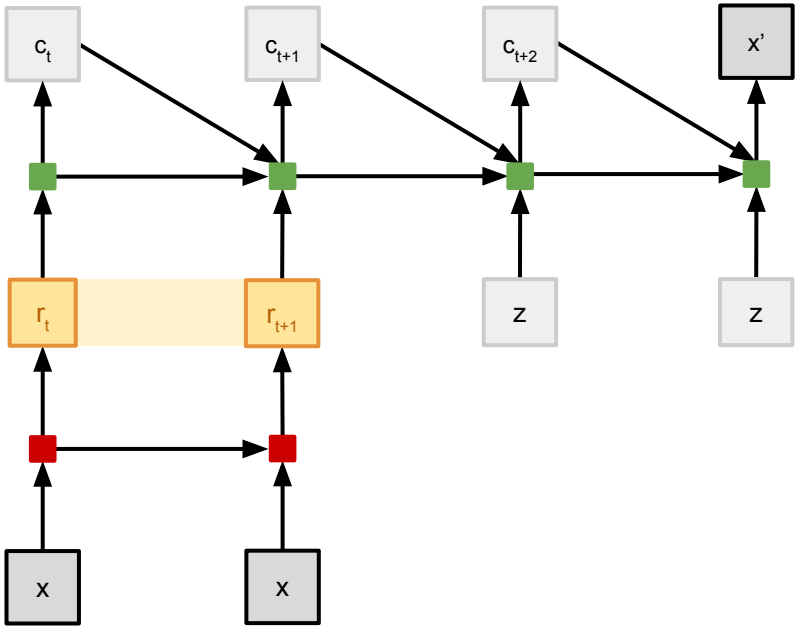
Generation RNN

Representation RNN

Sequential Autoencoders

Want to learn more?

Towards Conceptual Compression, Gregor et al, NeurIPS (2016)



Generation RNN

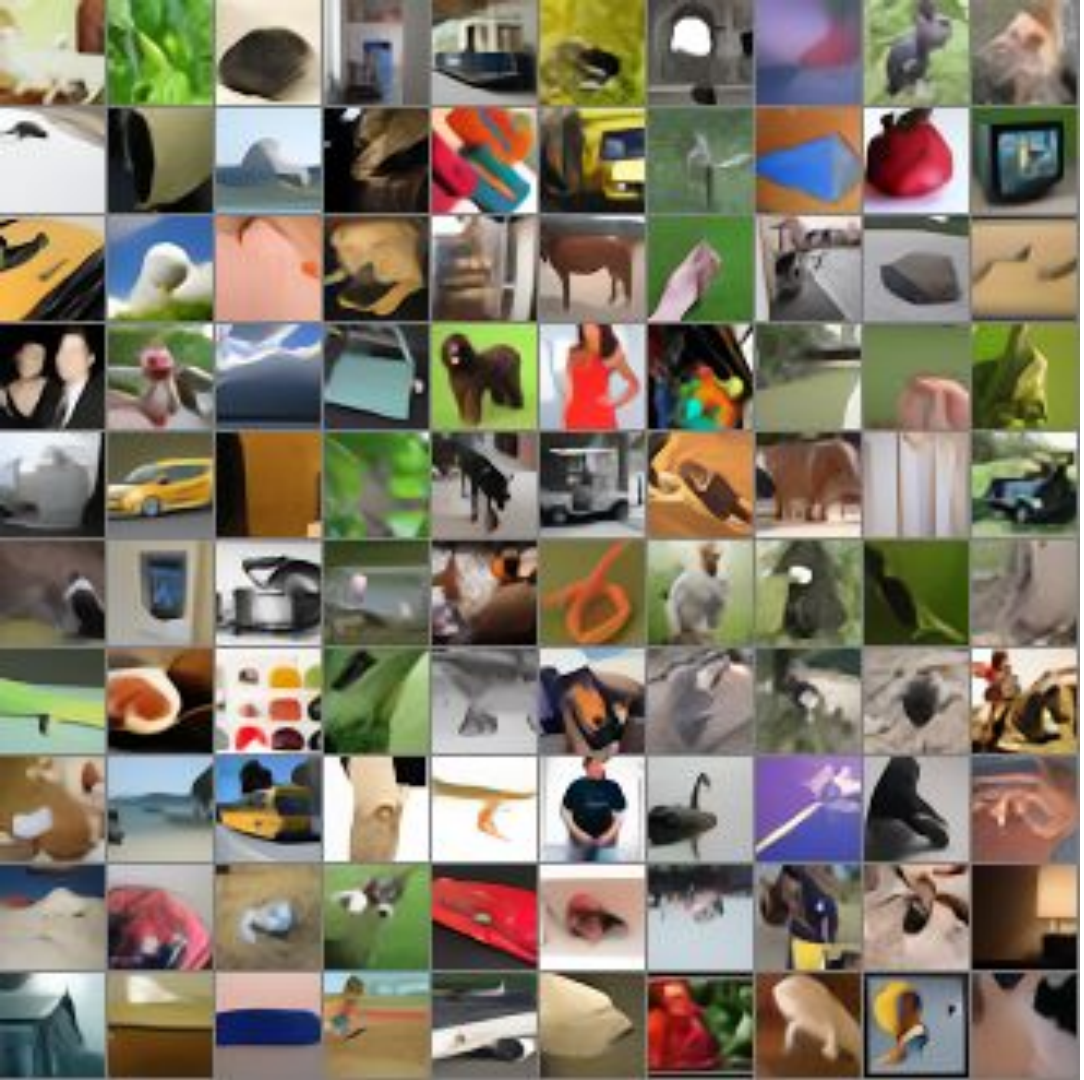
Representation RNN

76 bits

Original raw image:
24576 bits

112 bits

221 bits



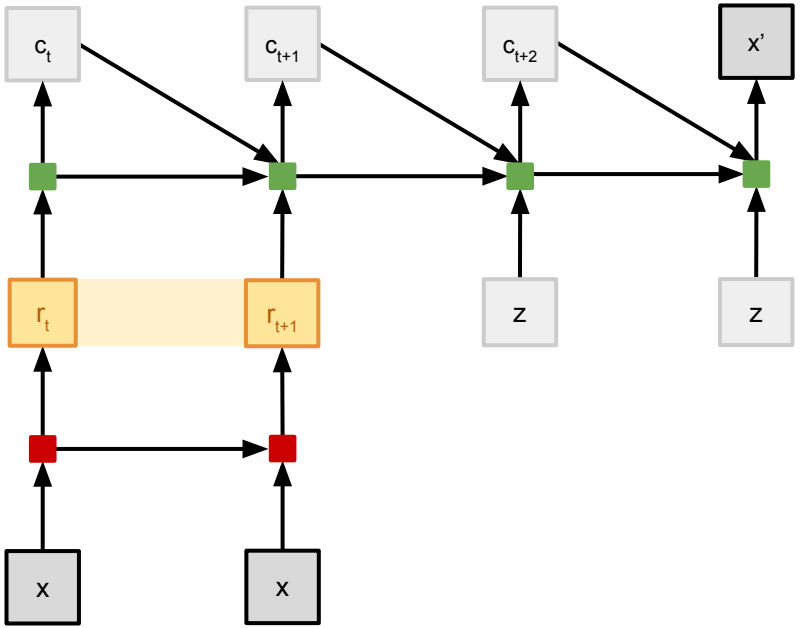
380 bits

2364 bits

Sequential Autoencoders

Want to learn more?

Towards Conceptual Compression, Gregor et al, NeurIPS (2016)



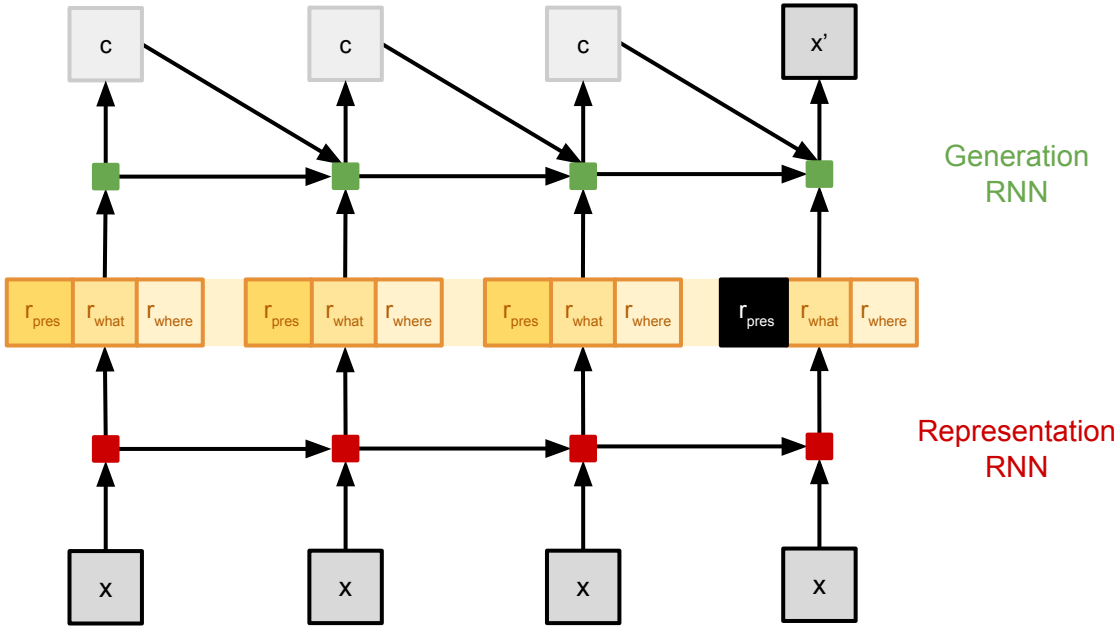
Generation RNN

Representation RNN

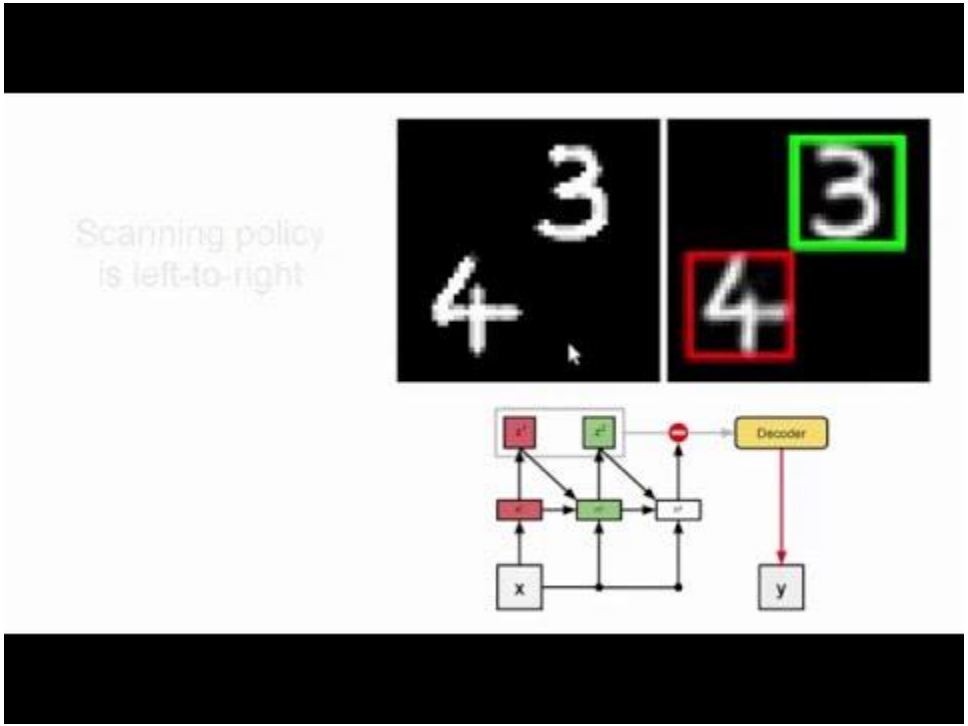
Variable Length, Interpretable, Sequential Autoencoders

Want to learn more?

Attend, Infer, Repeat, Eslami et al, NeurIPS (2016)



Variable Length, Interpretable, Sequential Autoencoders



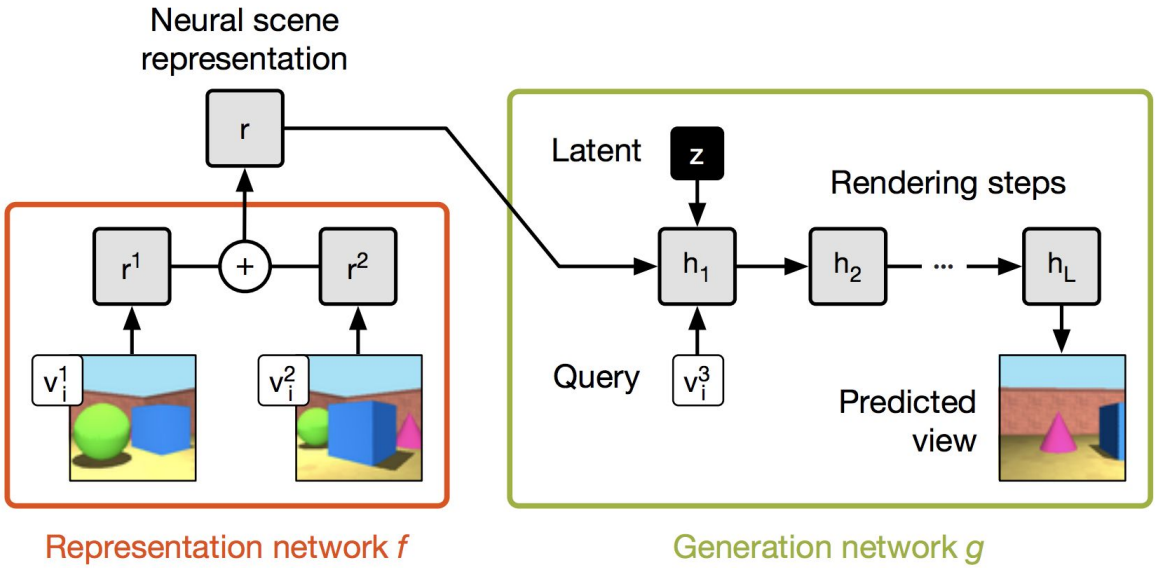
Want to learn more?

Attend, Infer, Repeat, Eslami et al, NeurIPS (2016)

Generative Query Networks

Want to learn more?

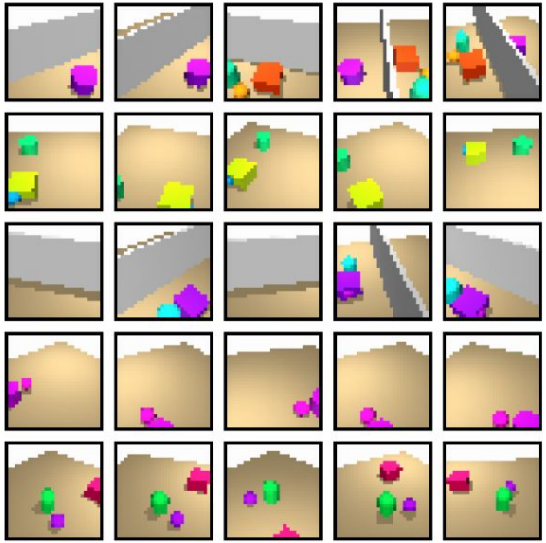
Neural scene representation and rendering, Eslami et al, Science (2018)



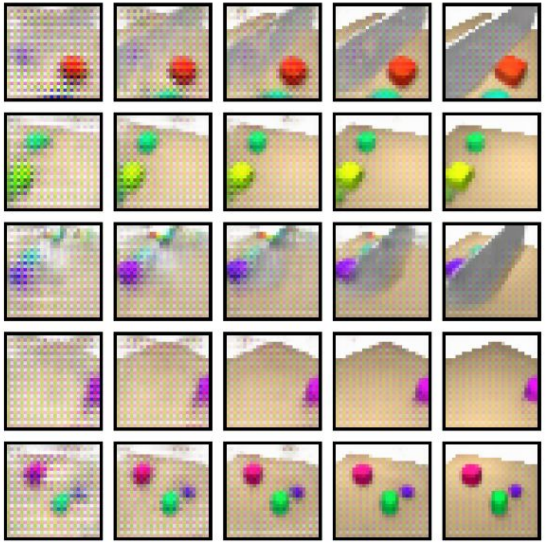
GQN: Accurate generation

Want to learn more?

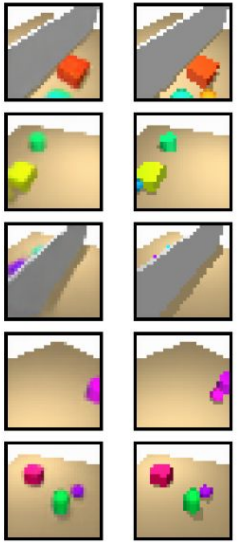
Neural scene representation and rendering, Eslami et al, Science (2018)



Observations



Generation steps

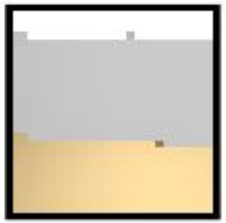


Pred Truth

GQN: Capturing uncertainty

Want to learn more?

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



Observation

Samples

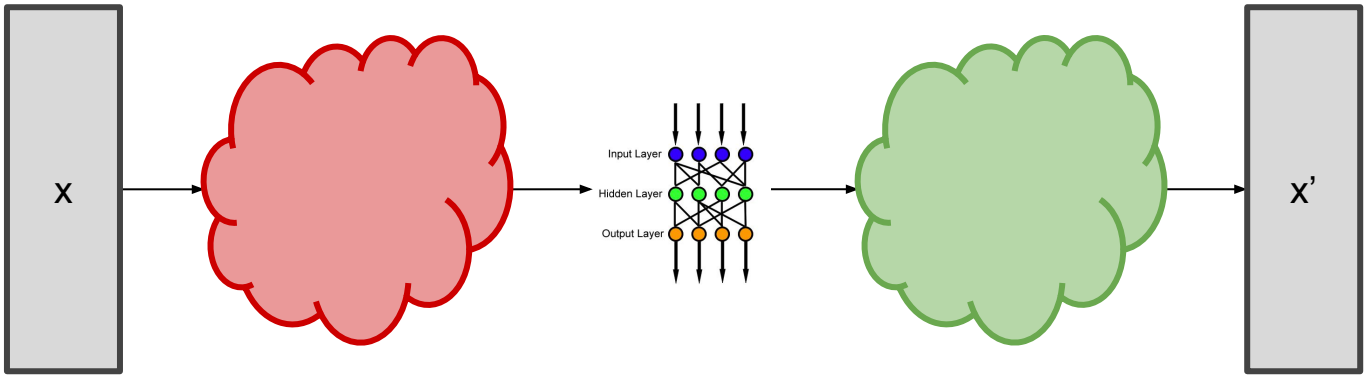
observations

ground truth | neural rendering | neural rendering | map

NeRF

Want to learn more?

NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, Mildenhall (2020)



Input

Inference Optimisation Algorithm

Neural Network Representation

Forward Propagation

Output

NeRF

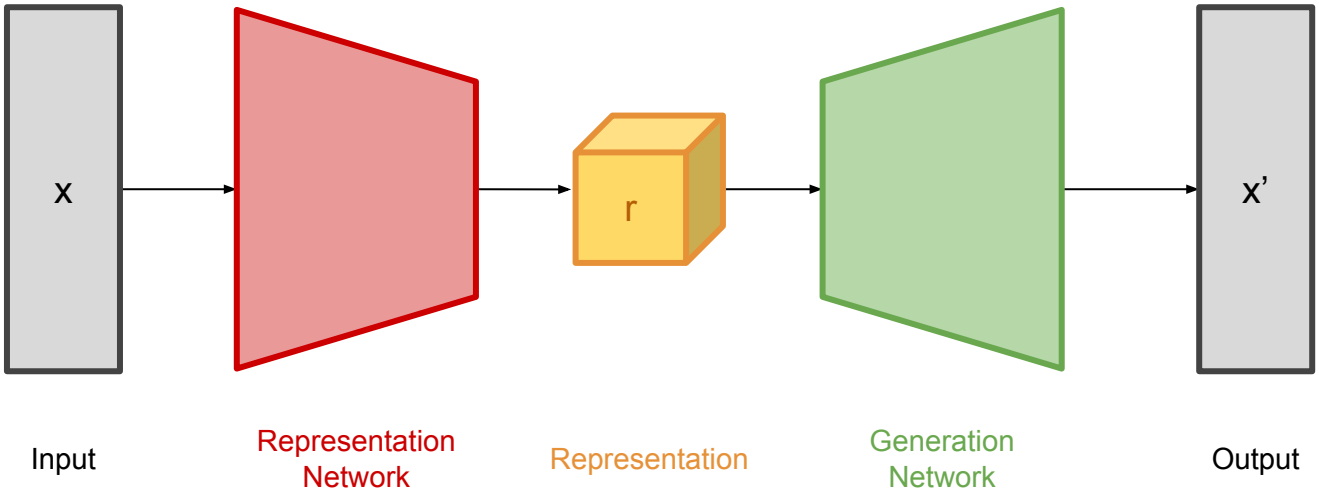
Want to learn more?

NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, Mildenhall (2020)

Voxel Autoencoders

Want to learn more?

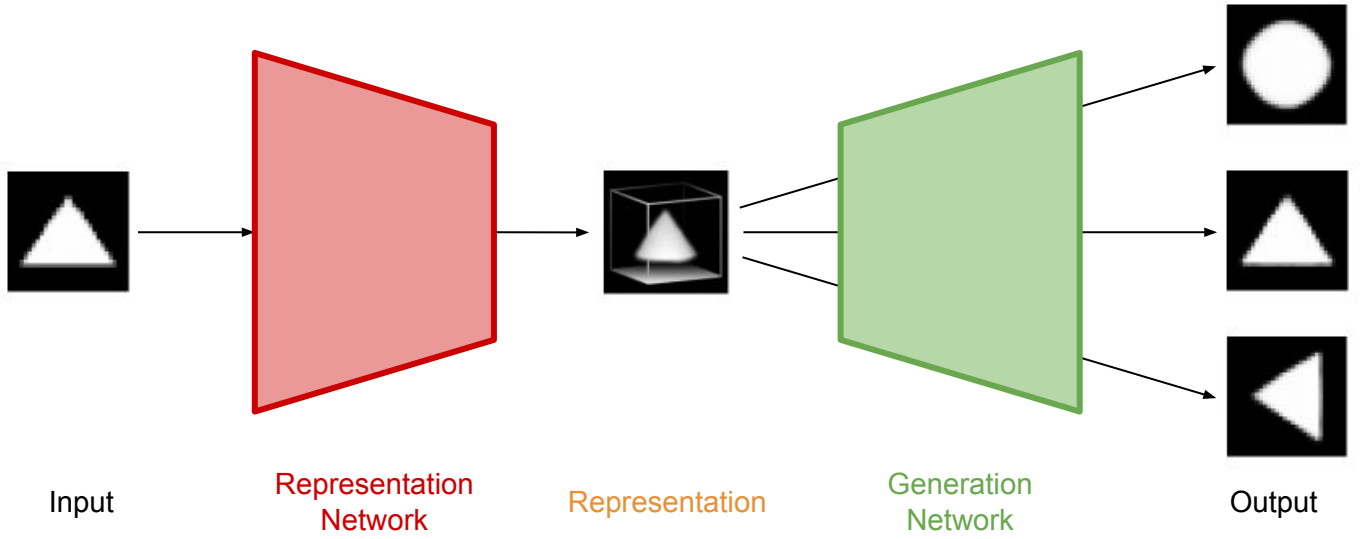
Unsupervised Learning of 3D Structure from Images, Rezende et al, NeurIPS (2016)



Voxel Autoencoders

Want to learn more?

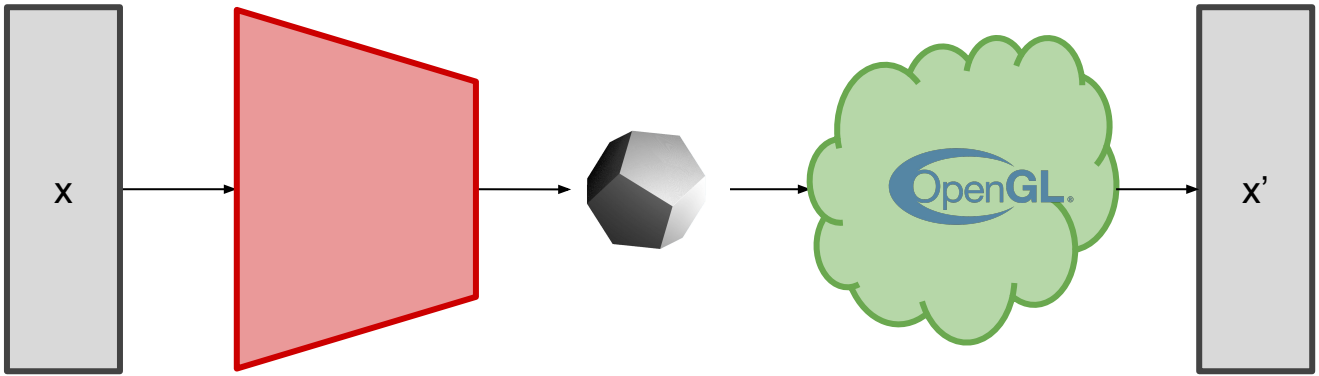
Unsupervised Learning of 3D Structure from Images, Rezende et al, NeurIPS (2016)



Mesh Autoencoders

Want to learn more?

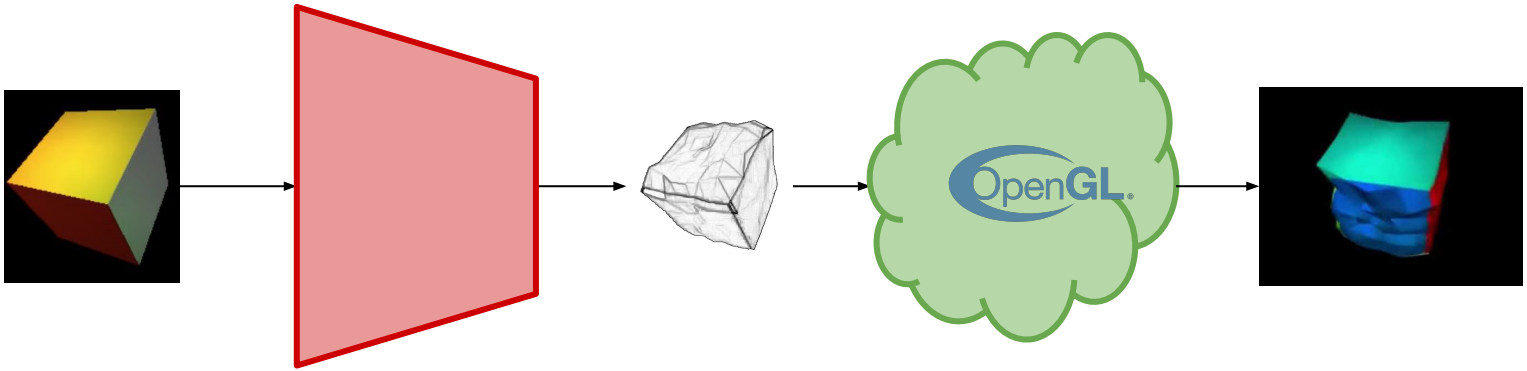
Unsupervised Learning of 3D Structure from Images, Rezende et al, NeurIPS (2016)



Mesh Autoencoders

Want to learn more?

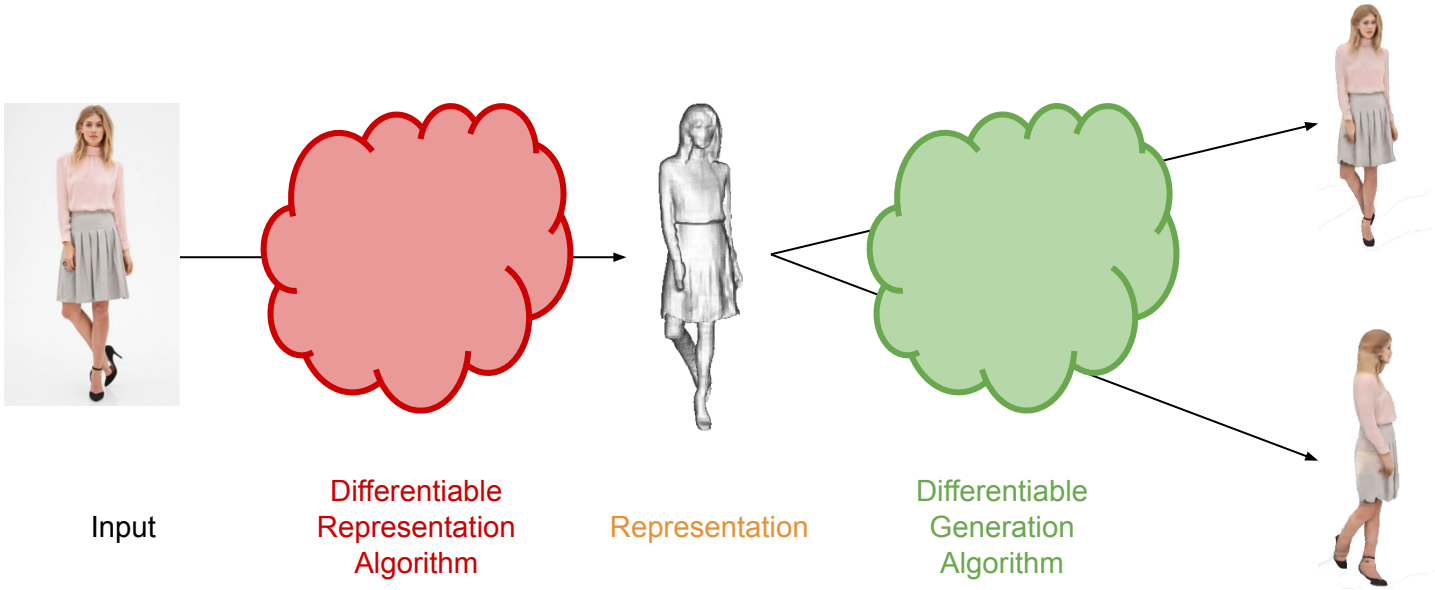
Unsupervised Learning of 3D Structure from Images, Rezende et al, NeurIPS (2016)



Implicit Function Autoencoders

Want to learn more?

PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization, Saito et al, (2019)



Beyond likelihood-based

Discriminators / Contrastive Networks

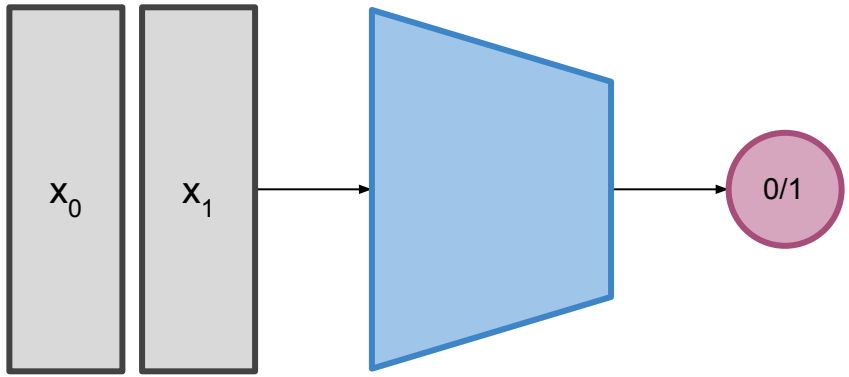


Diagram inspired by Lilian Weng, https://lilianweng.github.io

Generative adversarial networks

Want to learn more?

Generative adversarial networks. Goodfellow, et al. NeurIPS (2014)

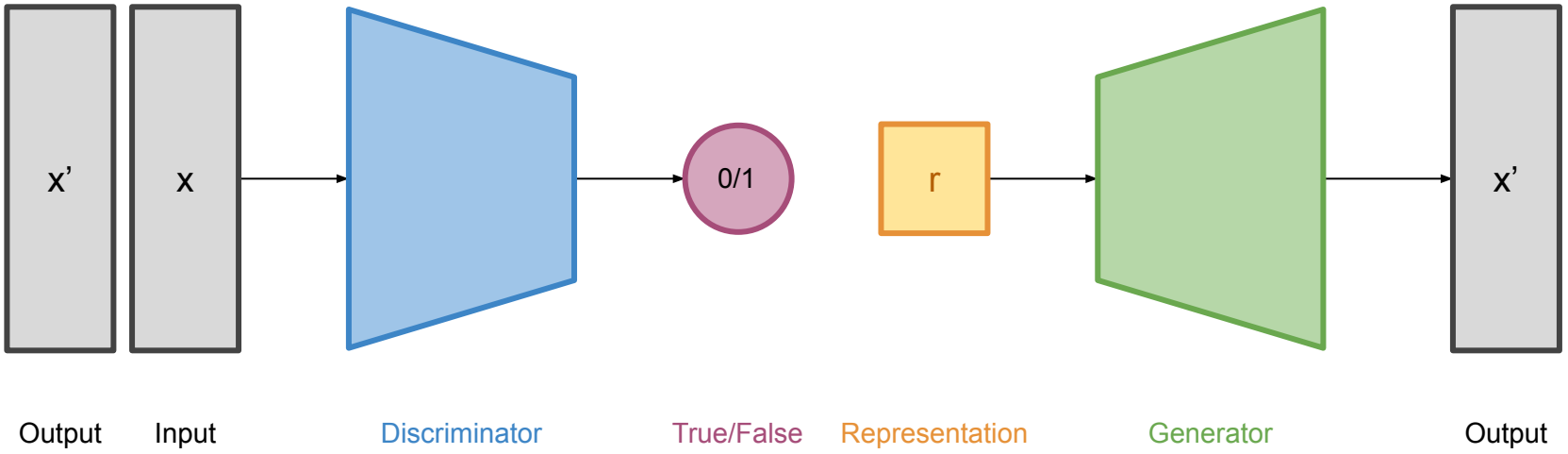


Diagram inspired by Lilian Weng, https://lilianweng.github.io

Generative adversarial networks

Want to learn more?

A Style-Based Generator for GANs, Karras et al (2018)

Large Scale GAN Training for High Fidelity Natural Image Synthesis, Brock et al (2018)

Generative adversarial networks

Want to learn more?

Generative adversarial networks. Goodfellow, et al. NeurIPS (2014)

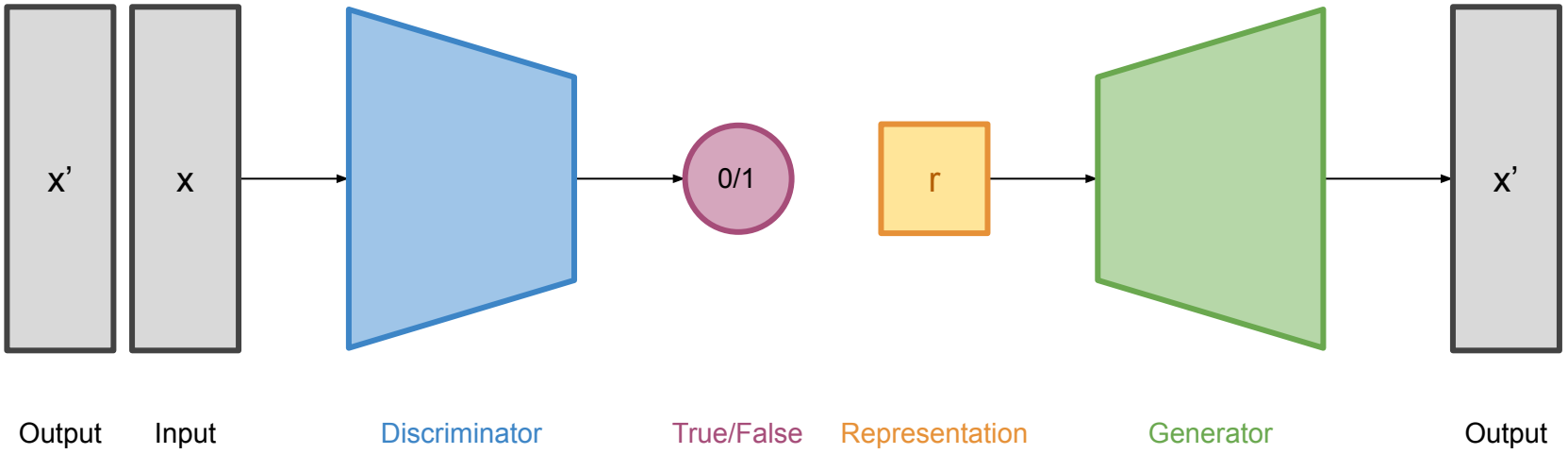
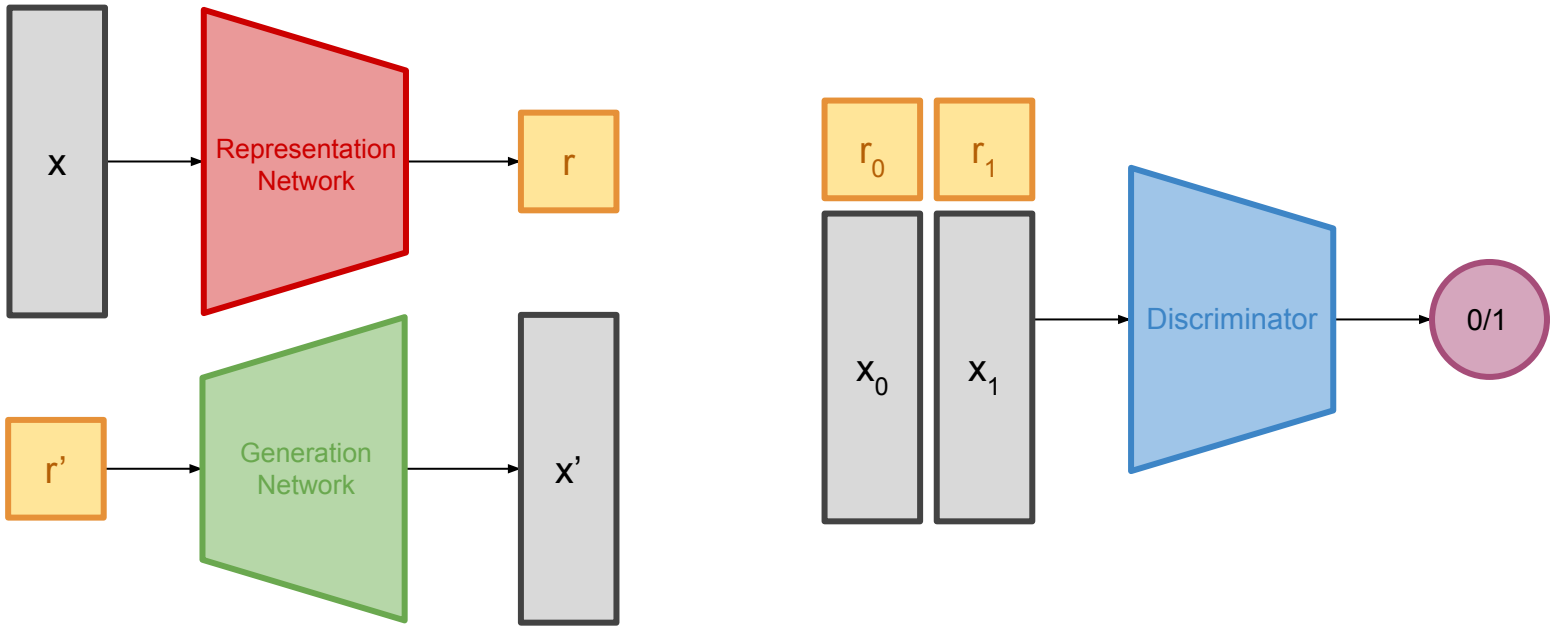


Diagram inspired by Lilian Weng, https://lilianweng.github.io

BiGAN

Want to learn more?

Adversarial Feature Learning, Donahue, et al. ICLR (2017)

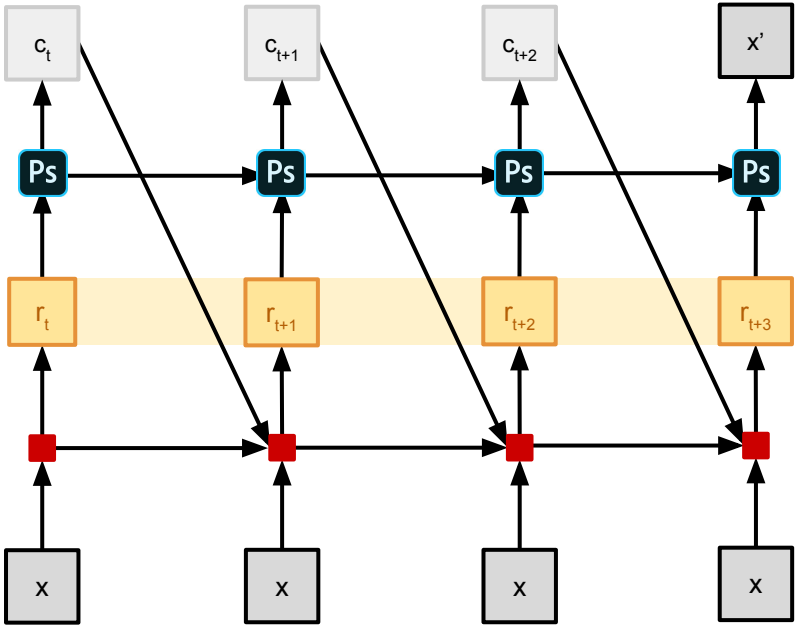


BigBiGAN

Want to learn more?

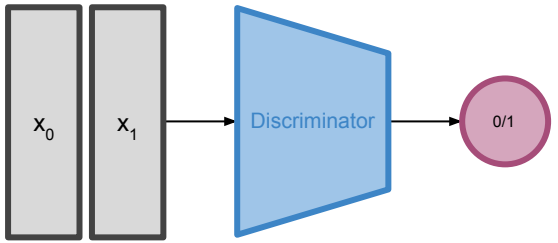
Large Scale Adversarial Representation Learning. Donahue, et al. NeurIPS (2019)

SPIRAL



Paintbrush Simulator

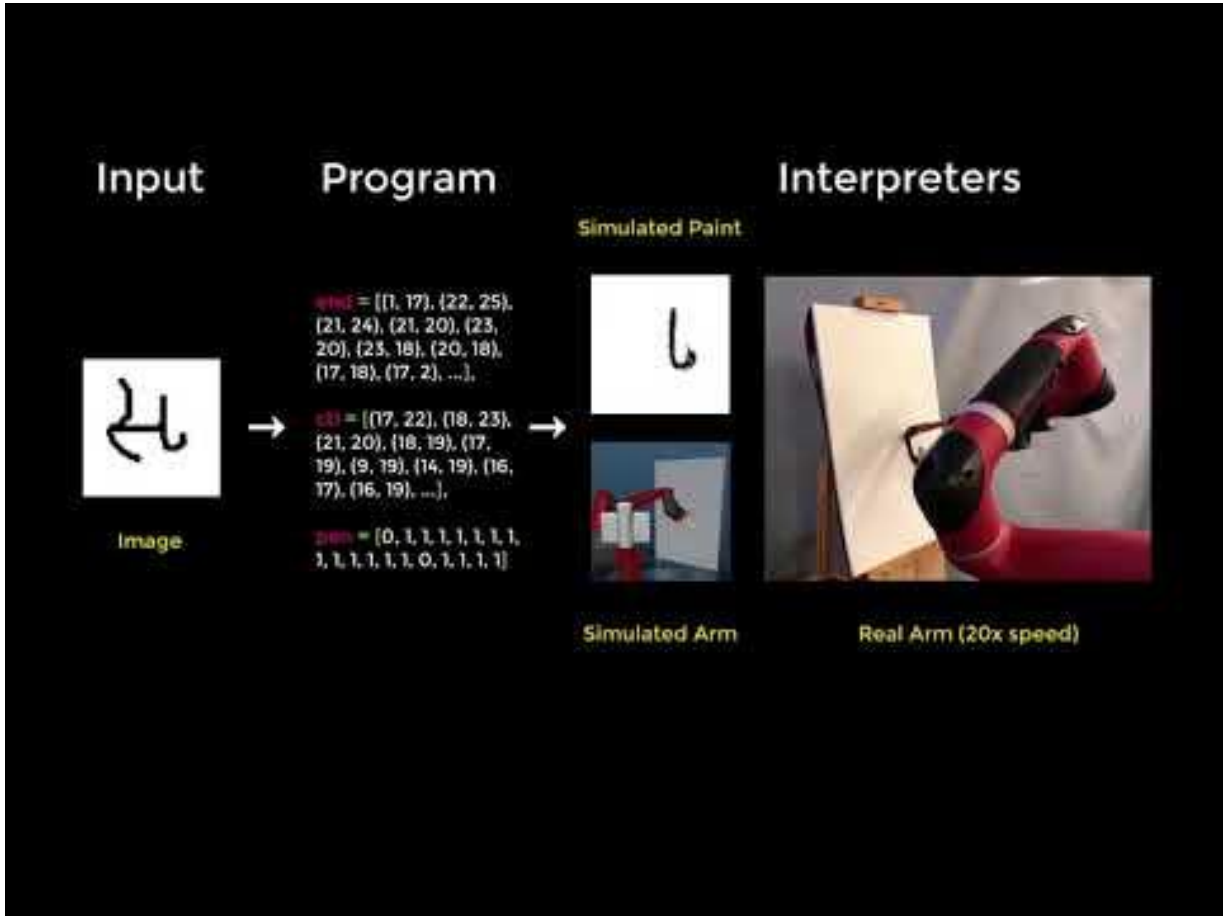
Representation RNN



Want to learn more?

Synthesizing Programs for Images using Reinforced Adversarial Learning, Ganin et al, ICML (2018)

Unsupervised Doodling and Painting with Improved SPIRAL, Mellor et al (2019)



Want to learn more?

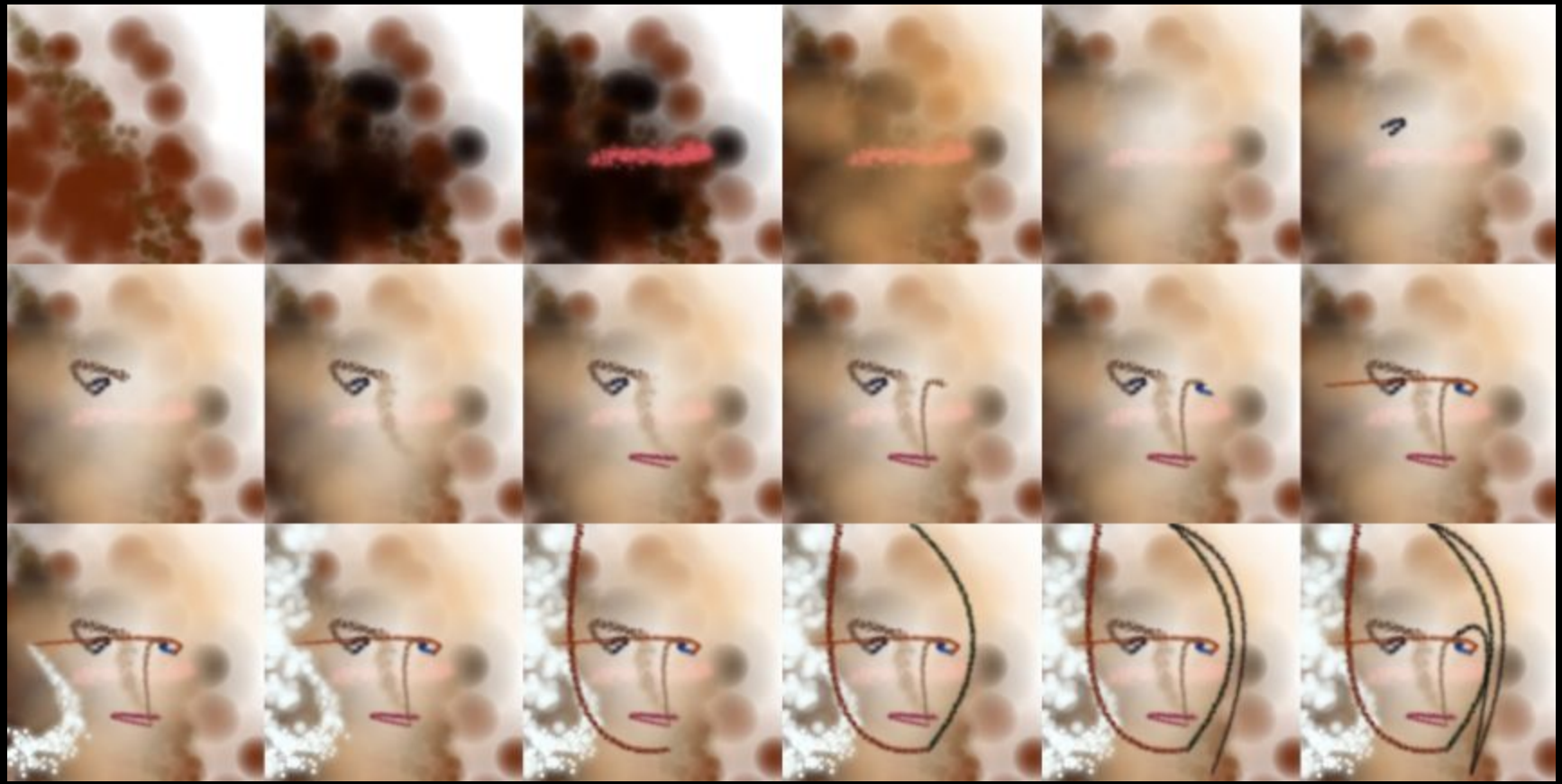
Synthesizing Programs for Images using Reinforced Adversarial Learning, Ganin et al, ICML (2018)

Unsupervised Doodling and Painting with Improved SPIRAL, Mellor et al (2019)

Want to learn more?

Synthesizing Programs for Images using Reinforced Adversarial Learning, Ganin et al, ICML (2018)

Unsupervised Doodling and Painting with Improved SPIRAL, Mellor et al (2019)

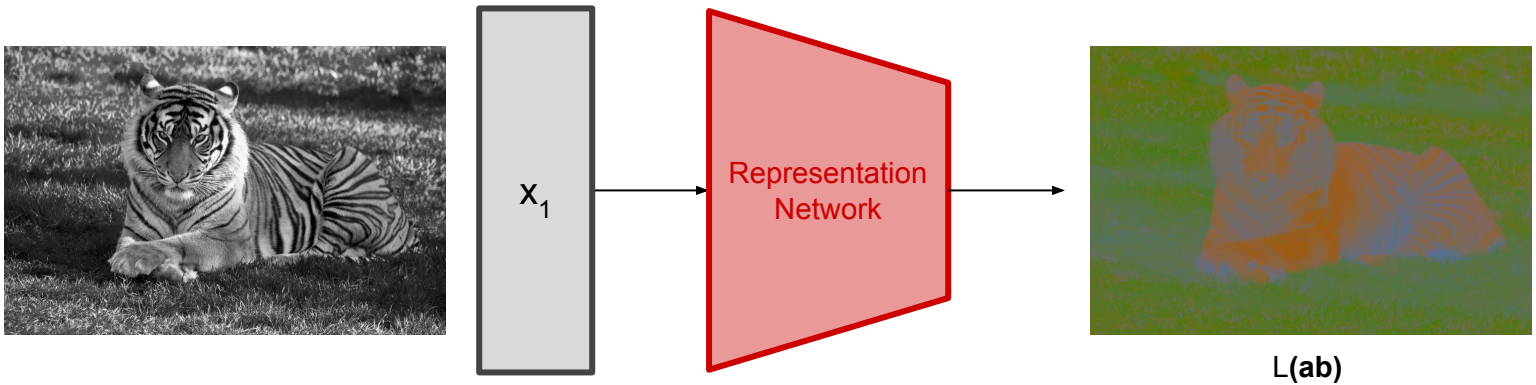


Beyond generative

Colorization

Want to learn more?

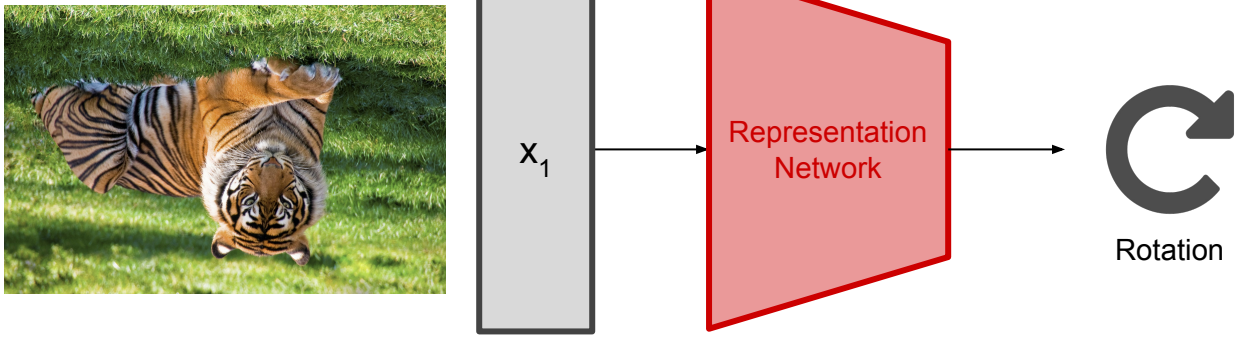
Colorization as a proxy task for visual understanding, Larsson et al, CVPR (2017)



Rotation Prediction

Want to learn more?

Unsupervised Representation Learning by Predicting Image Rotations, Gidaris et al, ICLR (2018)

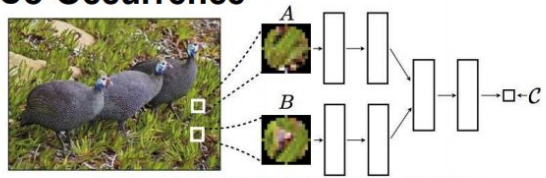


Self-supervised learning

Want to learn more?

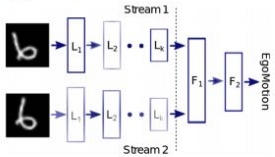
Self-Supervised Learning lecture, Andrew Zisserman, ICML (2018)

Co-Occurrence



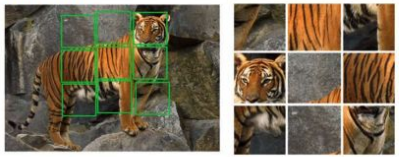
Isola *et al.* ICLR Workshop 2016.

Egomotion



Agrawal *et al.* ICCV 2015 Jayaraman *et al.* ICCV 2015

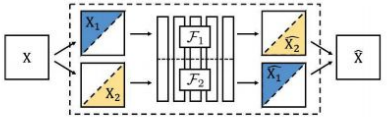
Context



Noroozi et al 2016

Pathak *et al.* CVPR 2016

Split-brain auto-encoders

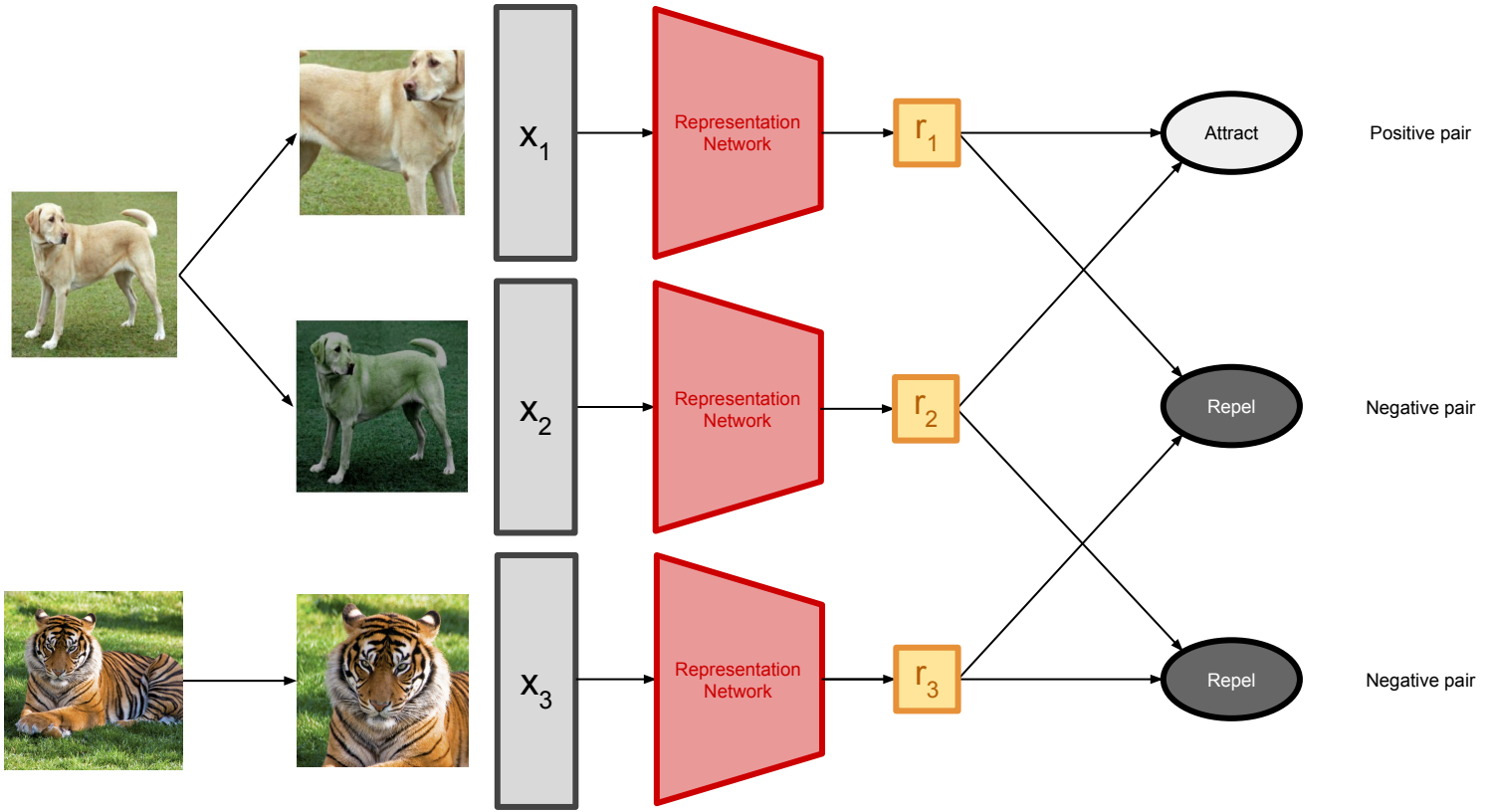


Zhang *et al.* CVPR 2017

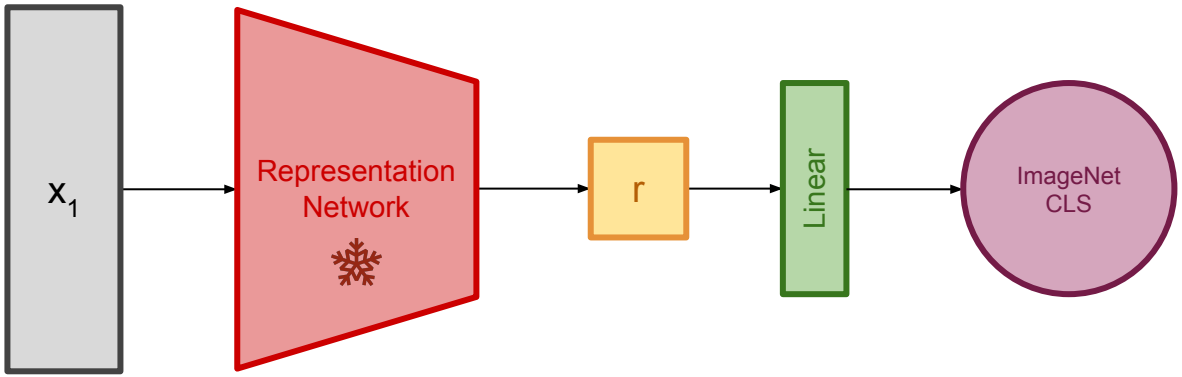
Contrastive learning

Want to learn more?

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



Evaluation: Linear separation



Contrastive learning

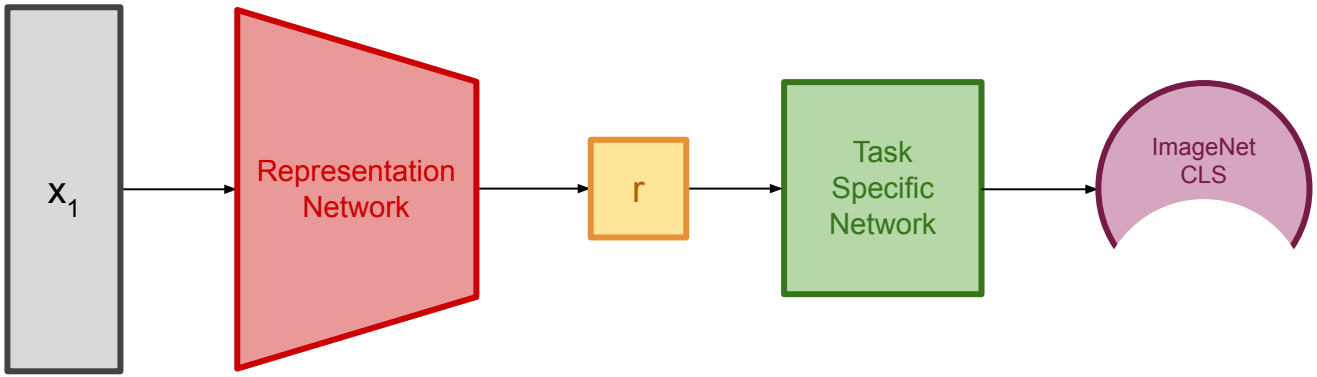
Want to learn more?

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

Method	Architecture	Param.	Top 1	Top 5
Methods using ResNet-50:				
Local Agg.	ResNet-50	24	60.2	-
MoCo	ResNet-50	24	60.6	-
PIRL	ResNet-50	24	63.6	-
CPC v2	ResNet-50	24	63.8	85.3
SimCLR (ours)	ResNet-50	24	**69.3**	**89.0**
Methods using other architectures:				
Rotation	RevNet-50 (4×)	86	55.4	-
BigBiGAN	RevNet-50 (4×)	86	61.3	81.9
AMDIM	Custom-ResNet	626	68.1	-
CMC	ResNet-50 (2×)	188	68.4	88.2
MoCo	ResNet-50 (4×)	375	68.6	-
CPC v2	ResNet-161 (∗)	305	71.5	90.1
SimCLR (ours)	ResNet-50 (2×)	94	74.2	92.0
SimCLR (ours)	ResNet-50 (4×)	375	**76.5**	**93.2**

Table 6. ImageNet accuracies of linear classifiers trained on representations learned with different self-supervised methods.

Evaluation: Data efficiency



Data efficient representation learning

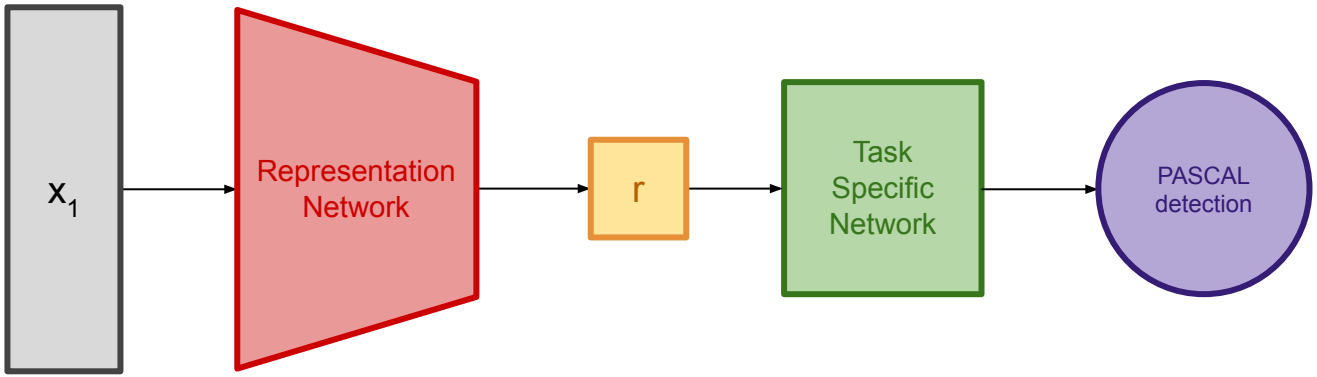
Want to learn more?

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

Method	Architecture	Label fraction 1%	10%
		Top 5	
Supervised baseline	ResNet-50	48.4	80.4
Methods using other label-propagation:			
Pseudo-label	ResNet-50	51.6	82.4
VAT+Entropy Min.	ResNet-50	47.0	83.4
UDA (w. RandAug)	ResNet-50	-	88.5
FixMatch (w. RandAug)	ResNet-50	-	89.1
S4L (Rot+VAT+En. M.)	ResNet-50 (4×)	-	91.2
Methods using representation learning only:			
InstDisc	ResNet-50	39.2	77.4
BigBiGAN	RevNet-50 (4×)	55.2	78.8
PIRL	ResNet-50	57.2	83.8
CPC v2	ResNet-161(*)	77.9	91.2
SimCLR (ours)	ResNet-50	75.5	87.8
SimCLR (ours)	ResNet-50 (2×)	83.0	91.2
SimCLR (ours)	ResNet-50 (4×)	**85.8**	**92.6**

Table 7. ImageNet accuracy of models trained with few labels.

Evaluation: Transfer learning



Transfer learning

Want to learn more?

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

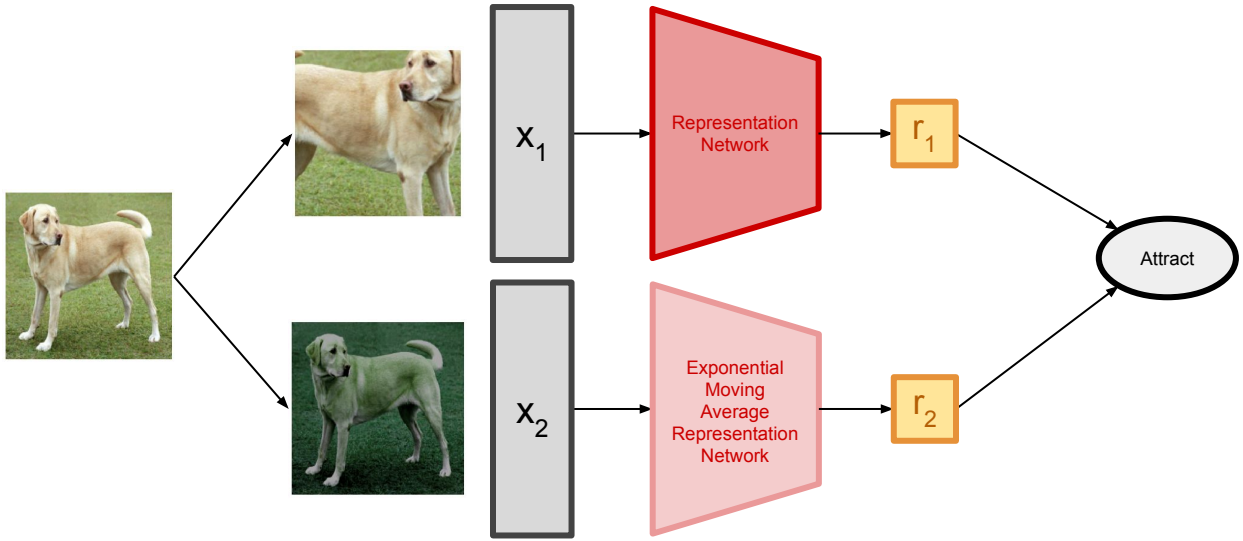
	Food	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech-101	Flowers
Linear evaluation:												
SimCLR (ours)	**76.9**	**95.3**	80.2	48.4	**65.9**	60.0	61.2	**84.2**	**78.9**	89.2	**93.9**	**95.0**
Supervised	75.2	**95.7**	**81.2**	**56.4**	64.9	**68.8**	**63.8**	83.8	**78.7**	**92.3**	**94.1**	94.2
Fine-tuned:												
SimCLR (ours)	**89.4**	**98.6**	**89.0**	**78.2**	**68.1**	**92.1**	**87.0**	**86.6**	**77.8**	92.1	**94.1**	97.6
Supervised	88.7	98.3	**88.7**	**77.8**	67.0	91.4	**88.0**	86.5	**78.8**	**93.2**	**94.2**	**98.0**
Random init	88.3	96.0	81.9	**77.0**	53.7	91.3	84.8	69.4	64.1	82.7	72.5	92.5

Table 8. Comparison of transfer learning performance of our self-supervised approach with supervised baselines across 12 natural image classification datasets, for ResNet-50 $(4\times)$ models pretrained on ImageNet. Results not significantly worse than the best ($p > 0.05$, permutation test) are shown in bold. See Appendix B.8 for experimental details and results with standard ResNet-50.

Bootstrap Your Own Latent

Want to learn more?

Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning, Grill et al, arxiv (2020)



Bootstrap Your Own Latent

Want to learn more?

Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning, Grill et al, arxiv (2020)

Method	Architecture	Param.	Top-1	Top-5
SimCLR [8]	ResNet-50 (2×)	94M	74.2	92.0
CMC [11]	ResNet-50 (2×)	94M	70.6	89.7
BYOL (ours)	ResNet-50 (2×)	94M	77.4	93.6
CPC v2 [29]	ResNet-161	305M	71.5	90.1
MoCo [9]	ResNet-50 (4×)	375M	68.6	-
SimCLR [8]	ResNet-50 (4×)	375M	76.5	93.2
BYOL	ResNet-50 (4×)	375M	78.6	94.2
BYOL	ResNet-200 (2×)	250M	**79.6**	**94.8**

Surprising return of likelihood-based models

Attention-based models

iGPT

Want to learn more?

Generative Pretraining from Pixels, Chen et al, ICML (2020)

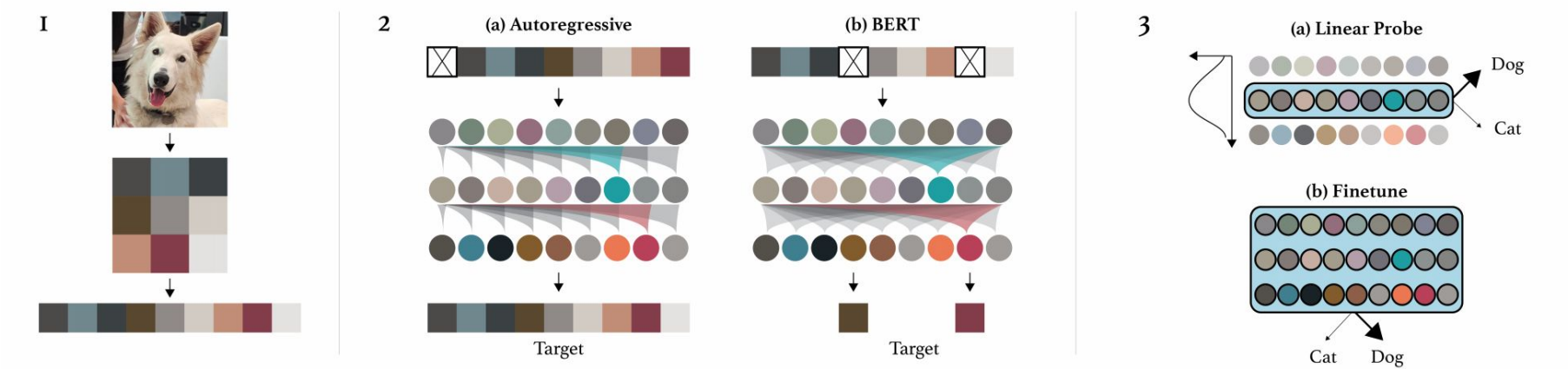
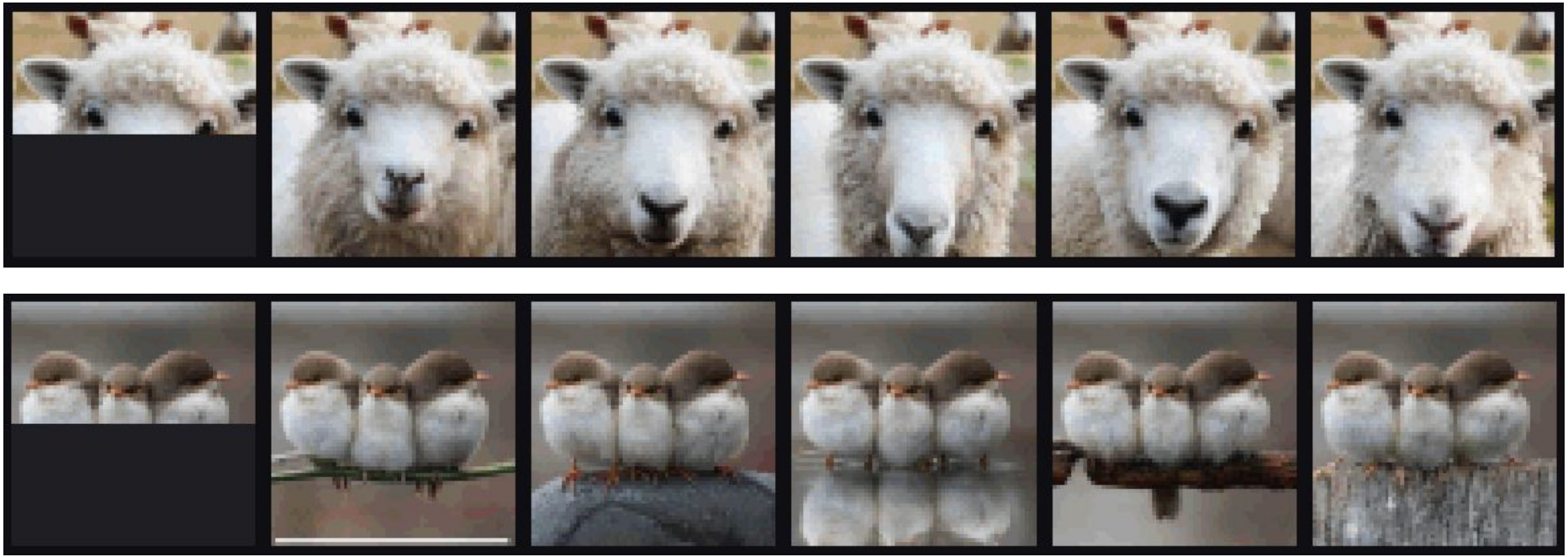


Figure 1. An overview of our approach. First, we pre-process raw images by resizing to a low resolution and reshaping into a 1D sequence. We then chose one of two pre-training objectives, auto-regressive next pixel prediction or masked pixel prediction. Finally, we evaluate the representations learned by these objectives with linear probes or fine-tuning.

iGPT

Want to learn more?

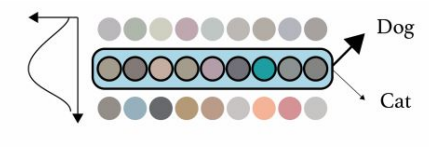
Generative Pretraining from Pixels, Chen et al, ICML (2020)

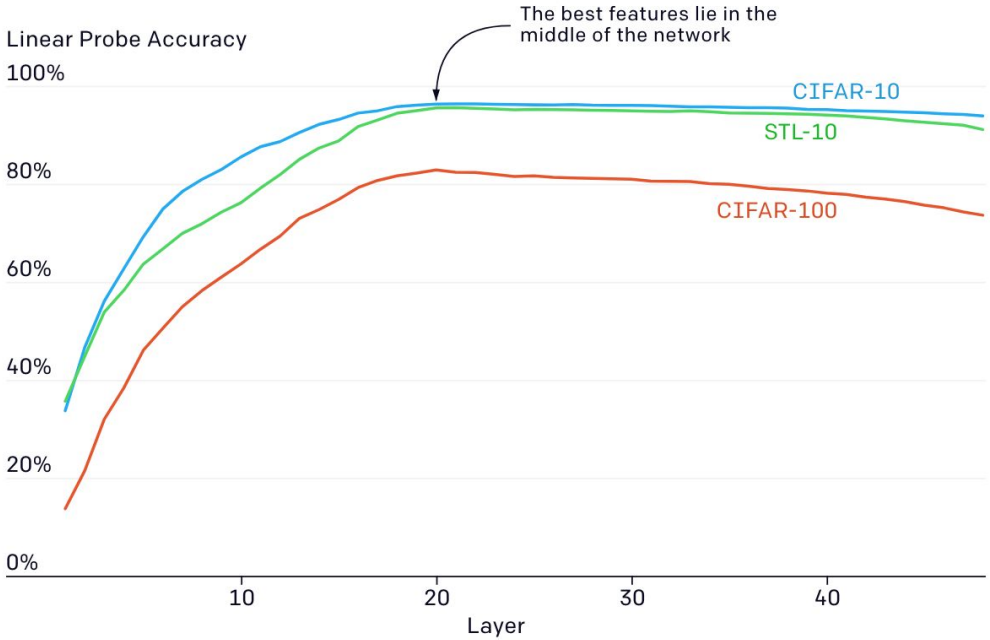


iGPT

Want to learn more?

Generative Pretraining from Pixels, Chen et al, ICML (2020)





Feature quality depends heavily on the layer we choose to evaluate. In contrast with supervised models, the best features for these generative models lie in the middle of the network.

iGPT

Want to learn more?

Generative Pretraining from Pixels, Chen et al, ICML (2020)

METHOD	INPUT RESOLUTION	FEATURES	PARAMETERS	ACCURACY
Rotation[53]	original	8192	86M	55.4
iGPT-L	32x32	1536	1362M	60.3
BigBiGAN[37]	original	8192	86M	61.3
iGPT-L	48x48	1536	1362M	65.2
AMDIM[13]	original	8192	626M	68.1
MoCo[24]	original	8192	375M	68.6
iGPT-XL	64x64	3072	6801M	68.7
SimCLR[12]	original	2048	24M	69.3
CPC v2[25]	original	8192	303M	71.5
iGPT-XL	64x64	3072 x 5	6801M	72.0
SimCLR	original	8192	375M	**76.5**

A comparison of linear probe accuracies between our models and state-of-the-art self-supervised models. We achieve competitive performance while training at much lower input resolutions, though our method requires more parameters and compute.

Summary

The representation learning problem is **under-specified**.

Two broad categories of approaches:

1. **Building in structure** or **inductive bias** to obtain the 'right' representations, e.g. structured autoencoders
2. **Training for proxy tasks** that can only be solved with the 'right' representations, e.g. contrastive learning

Current approaches seem to involve a trade-off between:

1. **Generality of the representation**, i.e. what range of downstream tasks the representation is good for
2. **Interpretability**, i.e. how much control we have on the representational space

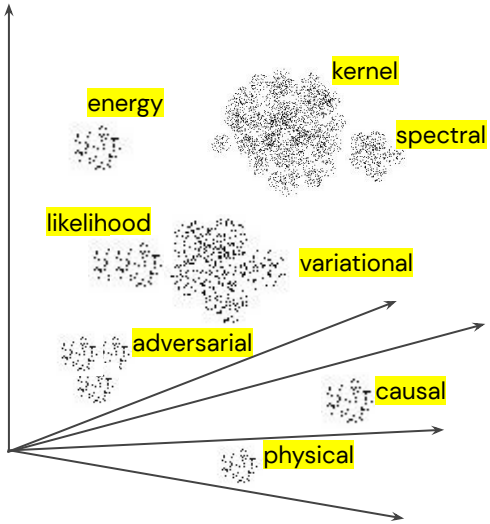
General representation learning without labels is **still largely unsolved**.

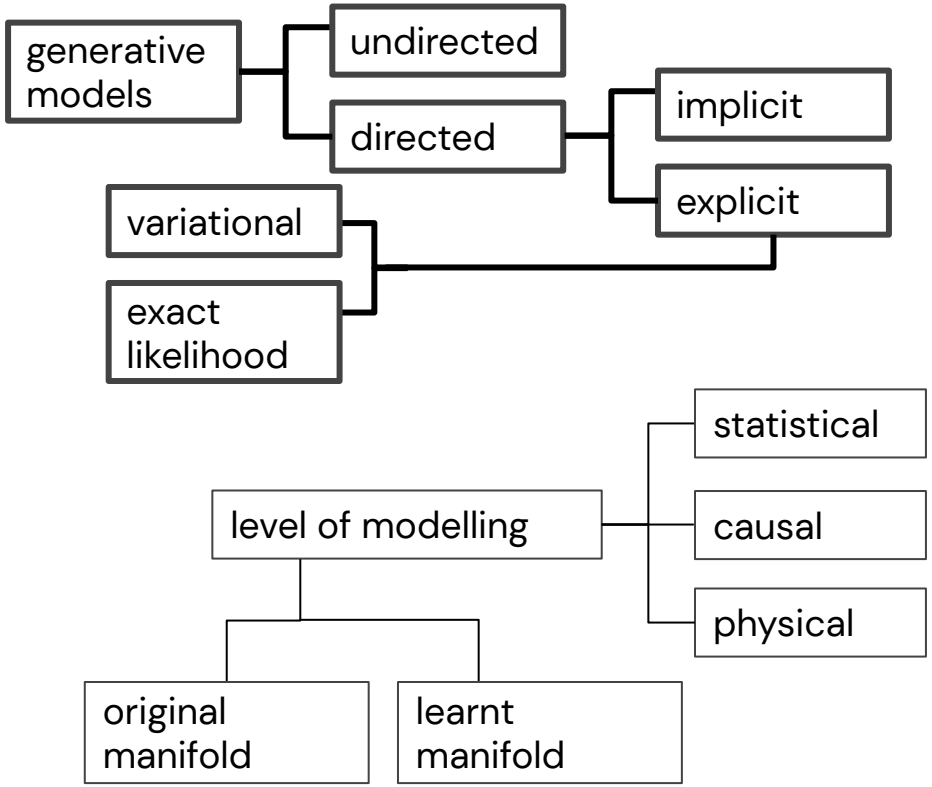
Recent advances, however, hold promise in finding increasingly general and useful representations.

DeepMind

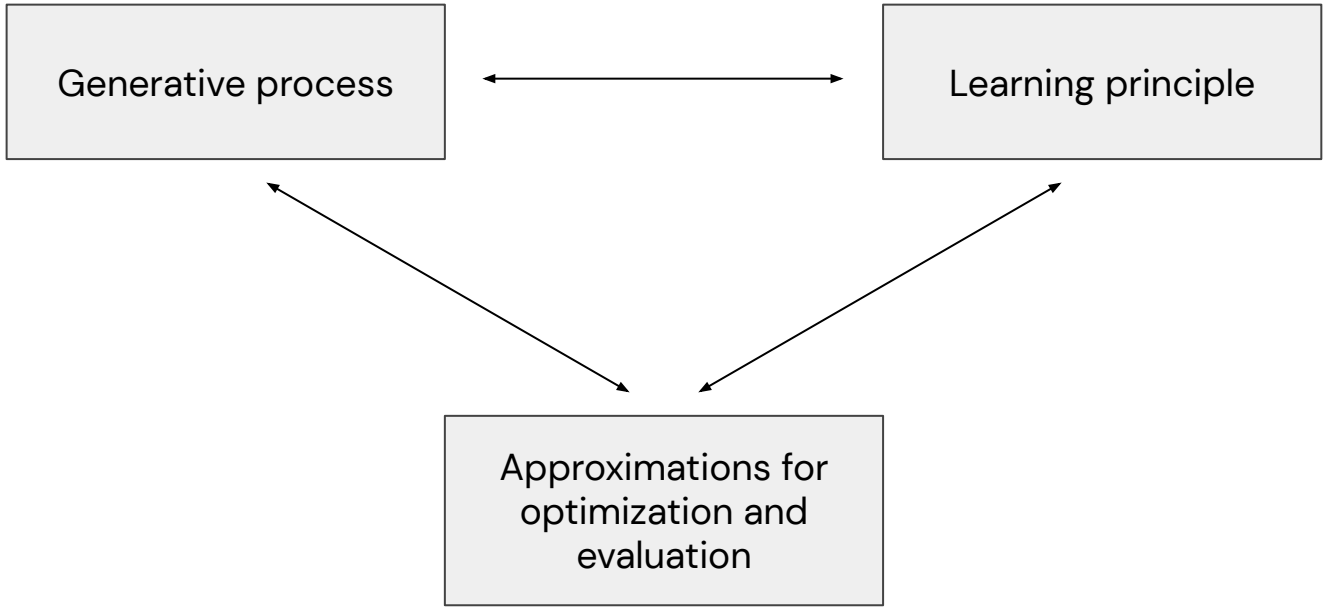
3 Landscape

Mapping out the landscape

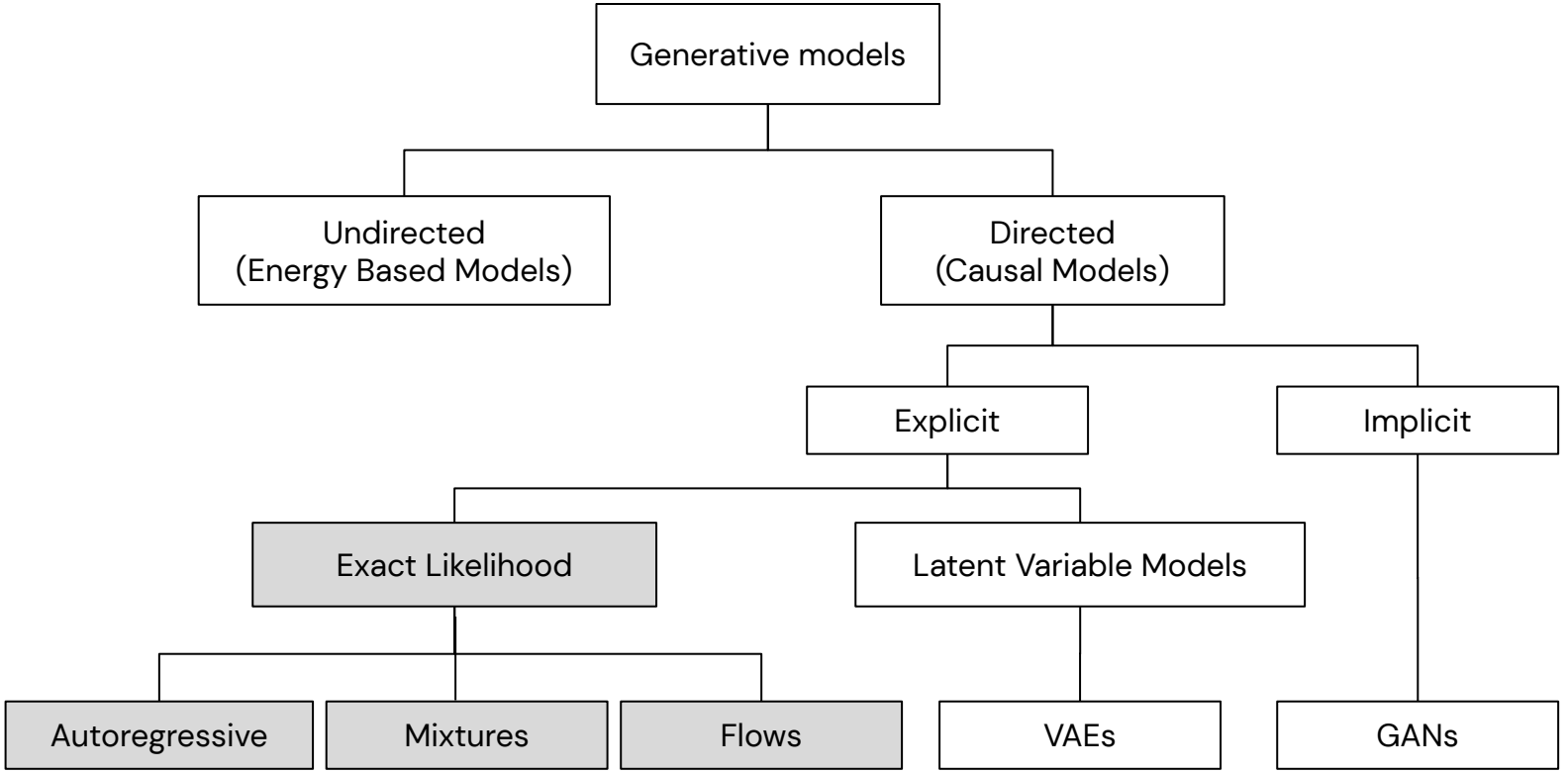




Foundations of generative models

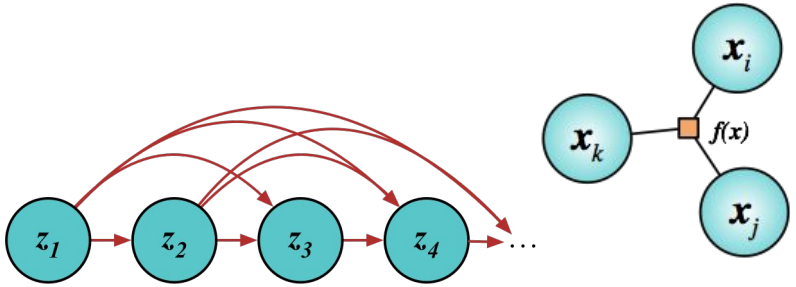


Mapping out the landscape of Generative Models



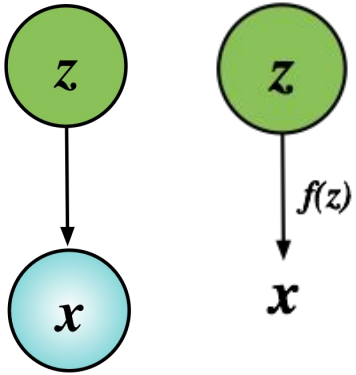
Types of Generative Models

Fully-observed models
Model observed data directly without introducing any new unobserved local variables.

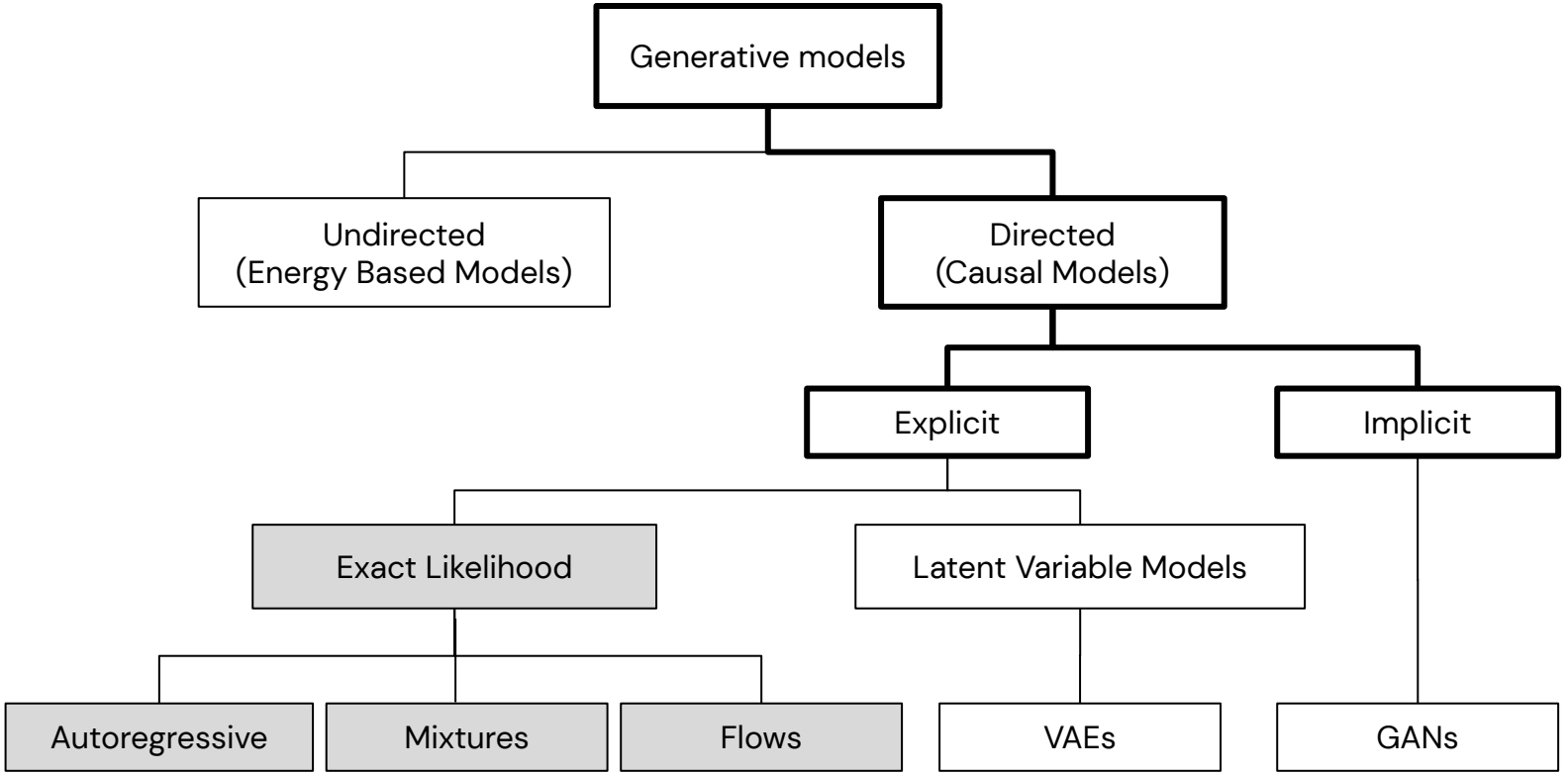


Latent Variable Models
Introduce an unobserved random variable for every observed data point to explain hidden causes.

- Prescribed models: Use observer likelihoods and assume observation noise.
- Implicit models: Likelihood-free models.



Mapping out the landscape of Generative Models



Foundations: density estimation & divergences

- Given an empirical density **p(x)** represented by a collection of iid samples **$\{x_1, ..., x_N\}$**
- Our goal is to modify the parameters **θ** of a parametric density **$q_\theta(x)$** so that it gets "closer" to **p(x)**
- But what means "closer"?

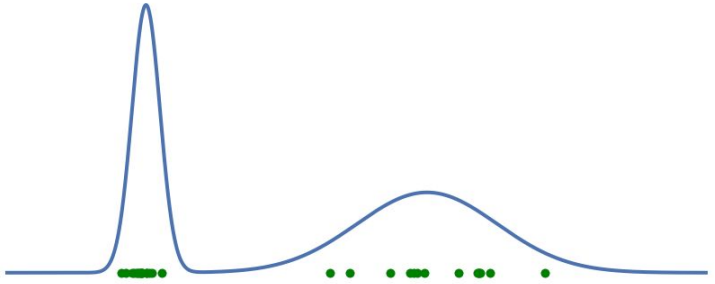


Image credit: Mihaela Rosca

Let $\mathcal{H}$ be the space of probability densities of interest (e.g. $\mathcal{H} = \{p : \mathbb{R} \to \mathbb{R}^+, ||p|| = 1\}$).
A *probability divergence* $\mathrm{D}(P;Q)$ is a map $\mathrm{D}: \mathcal{H}^2 \to \mathbb{R}$ such that:
For any P and Q, $\mathrm{D}(P;Q) \geq 0$
For any P and Q, $\mathrm{D}(P;Q) = 0 \Leftrightarrow P = Q$

Foundations: density estimation & divergences

Want to learn more?

Invariance, Rezende, Posts on ML, Math and Physics 2018
https://danilorezende.com/2018/07/12/short-notes-on-divergence-measures/

Each divergence will emphasize different aspects of the learned density

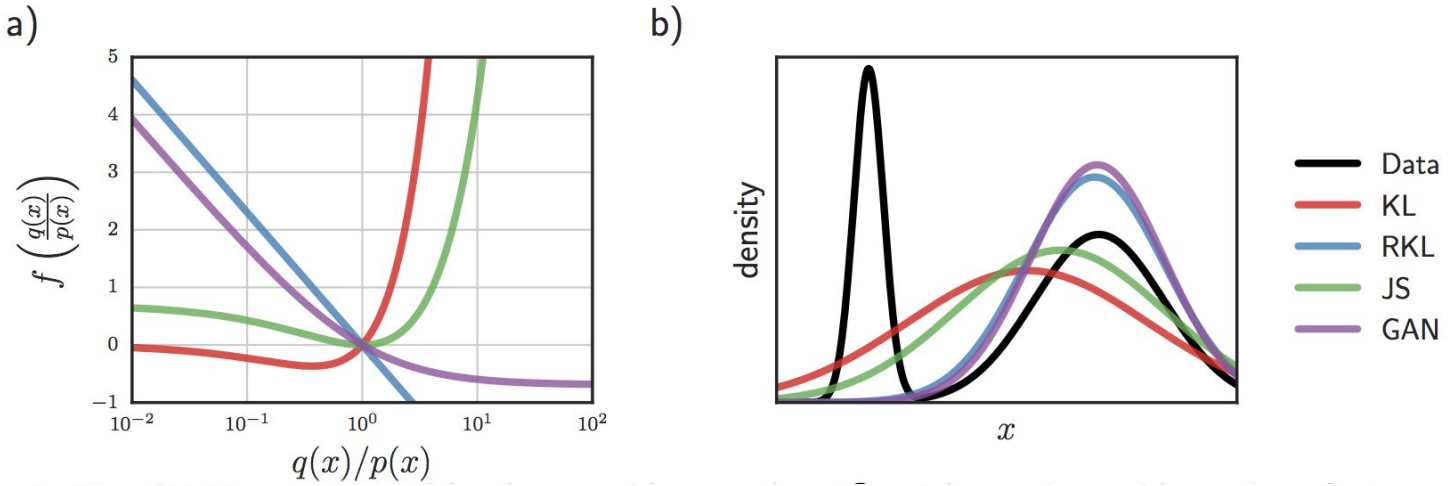
Name	Formula	Condition
f-divergences	$D(p;q) = \mathbb{E}_q[f(\frac{q}{p})]$	strictly convex f
Relative entropy (KL)	$D(p;q) = \mathbb{E}_q[\ln \frac{p}{q}]$	
Jensen-Shannon (JS)	$D(p;q) = \frac{1}{2}\mathrm{KL}(p;m) + \frac{1}{2}\mathrm{KL}(q;m)$	$m = \frac{1}{2}(p+q)$
Stein divergence	$D(p;q) = \sup_f \mathbb{E}_q[\nabla \ln pf + \nabla f]^2$	where $\int \nabla(pf) = 0$
Energy distance	$D(p;q) = \mathbb{E}[2\|x-y\| - \|x-x'\| - \|y-y'\|]$	$x, x' \sim p; y, y' \sim q$
Wasserstein distance	$D_\alpha(p;q) = [\inf_\rho \mathbb{E}_\rho[\|x-x'\|^\alpha]]^{\frac{1}{\alpha}}$	$\int dx\rho(x,x') = q(x')$ and $\int dx'\rho(x,x')$
Max-min dis. (MMD)	$D(p;q) = \sup_f(\mathbb{E}[f]_p - \mathbb{E}[f]_q)$	f continuous and bounded

Foundations: density estimation & divergences

Want to learn more?

Improved generator objectives for GANs, Poole et al, NeurIPS Workshop on Adversarial Training 2016

Each divergence will emphasize different aspects of the learned density



Foundations: density estimation & divergences

Want to learn more?

Estimating divergence functionals and the likelihood ratio by convex risk minimization, Nguyen, et al, IEEE Transactions on Information Theory 2010

A general bound on f-Divergences

Requires knowledge of **p(x)** and **q(x)**

$$D_f[p;q] = \mathbb{E}_q\left[f\left(\frac{p}{q}\right)\right] = \sup_{\phi \in L^2} \mathbb{E}_p[\phi] - \mathbb{E}_q[f^\star \circ \phi]$$

Only requires samples from **p** and **q** and evaluation of **ϕ(x)**

This observation is the key insight of GANs, allowing us to train models from which we can sample from but not evaluate its likelihood

Integral Probability Metrics

$$\mathcal{M}_f(p, q) = \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{p(x)}[f] - \mathbb{E}_{q_\theta(x)}[f] \right|$$

f sometimes referred to as a **test function, witness function or a critic.**

Many choices of ***f*** available: classifiers or functions in specified spaces.

$\|f\|_L < 1$
Wasserstein

$\|f\|_\infty < 1$
Total Variation

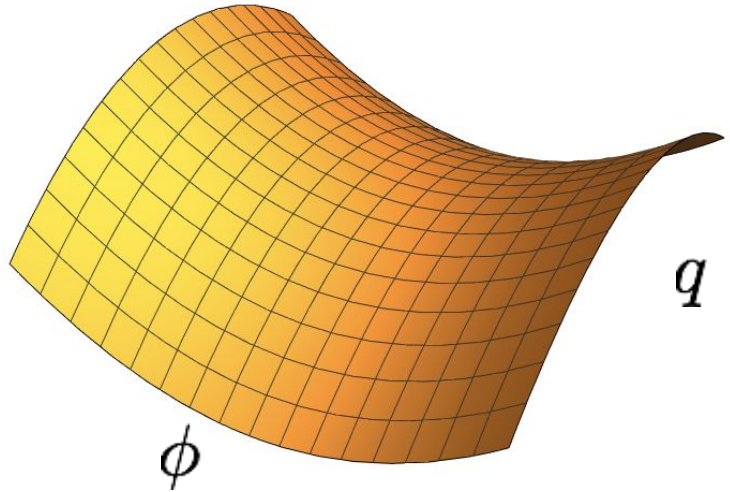
$\|f\|_{\mathcal{H}} < 1$
Max Mean Discrepancy

$\left\| \frac{df}{dx} \right\|_L < 1$
Cramer

Foundations: density estimation & divergences

A general bound on f-Divergences

$$\min_q D_f[p;q] = \min_q \max_\phi \mathbb{E}_p[\phi] - \mathbb{E}_q[f^\star \circ \phi]$$

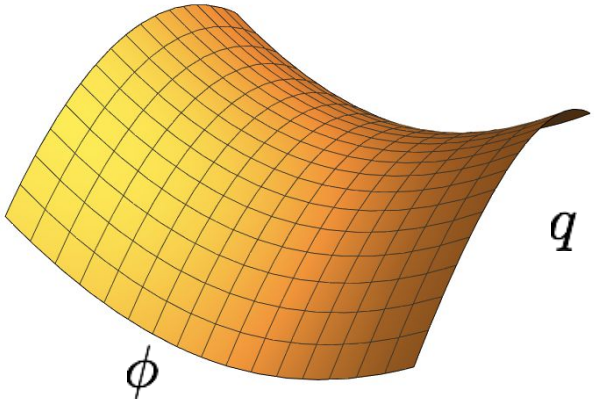


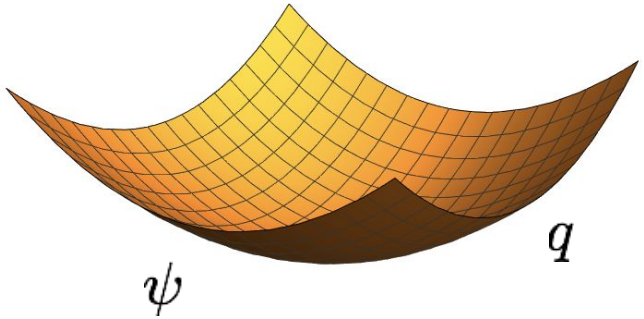
Foundations: density estimation & divergences

Upper and lower bounds on f-Divergences

$$f(u) = u \ln u$$

$$\mathbb{E}_p[\phi] - \mathbb{E}_q[f^\star \circ \phi] \leq D_f[p;q] \leq -\mathrm{ELBO}_\psi[p;q] + cst$$



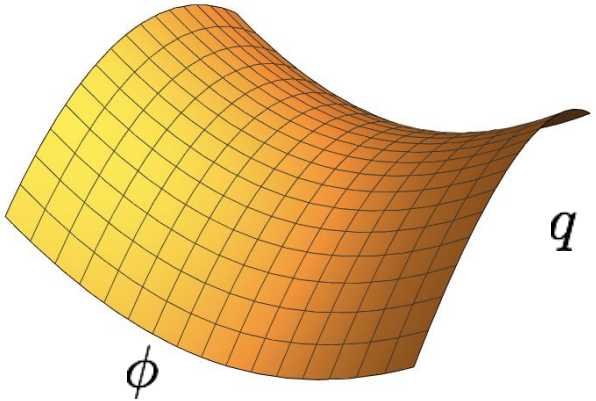


Foundations: density estimation & divergences

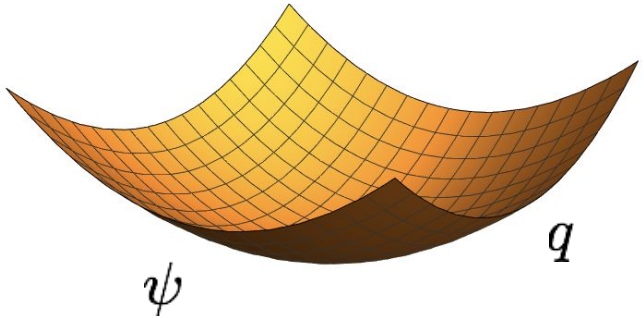
Upper and lower bounds on f-Divergences

$$f(u) = u \ln u$$

$$\mathbb{E}_p[\phi] - \mathbb{E}_q[f^\star \circ \phi] \leq D_f[p;q] \leq -\mathrm{ELBO}_\psi[p;q] + cst$$



GANs



VAEs

Foundations: density estimation & divergences

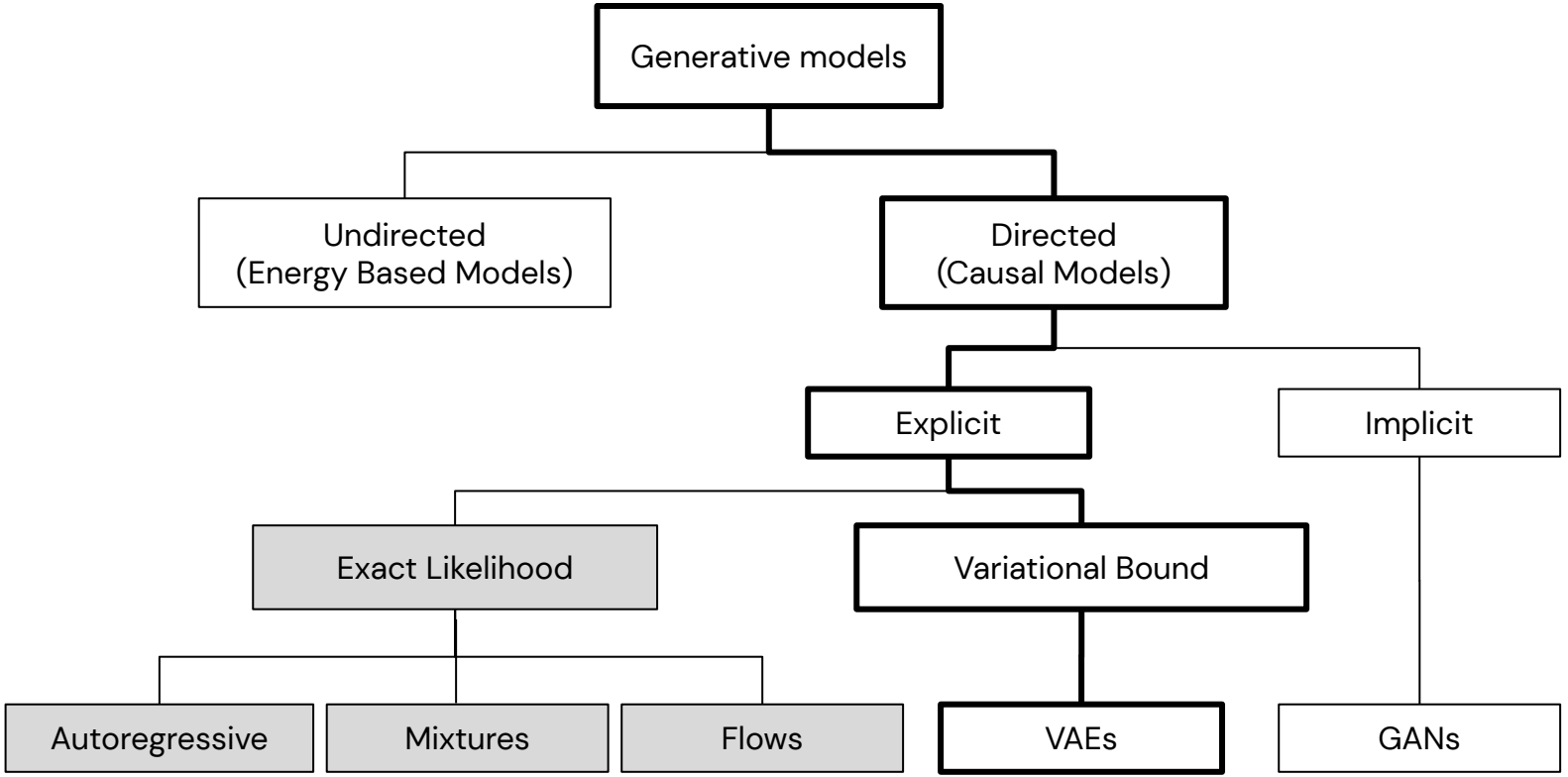
Want to learn more?

Improved generator objectives for GANs, Poole et al, NeurIPS Workshop on Adversarial Training 2016

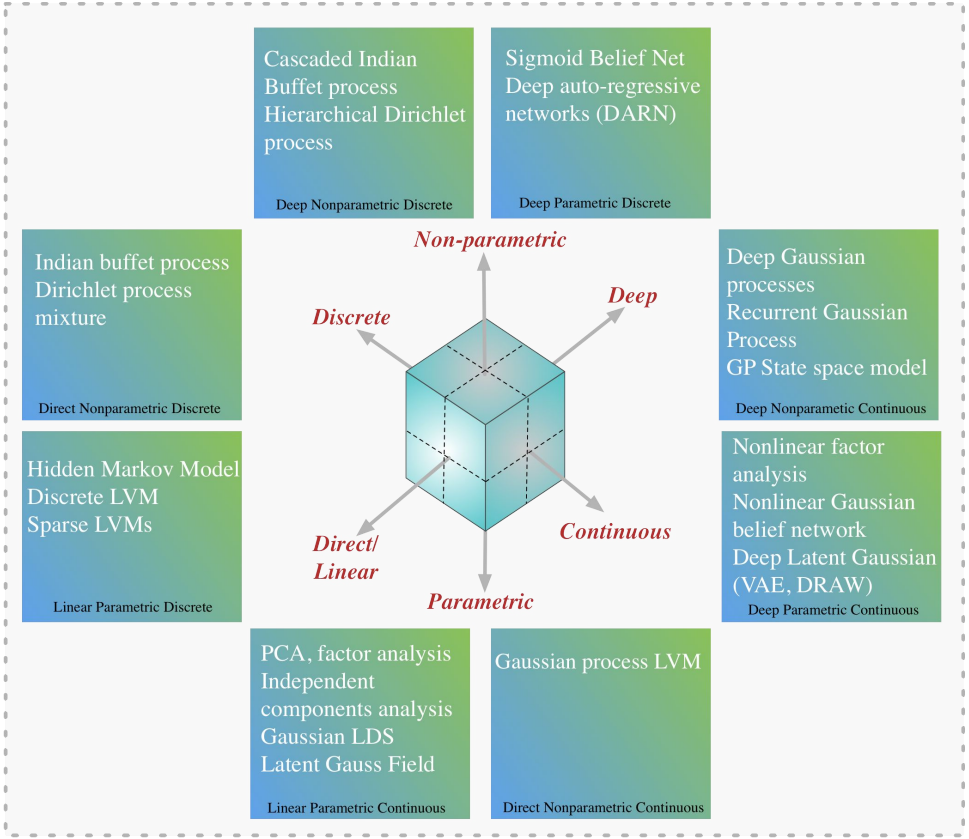
Many GAN models are particular cases of f-Divergences optimization

Name	Generator f-divergence (f_G)	Generator objective (minimized)
GAN-standard	$\log(1+\frac{1}{u})$	$\log\left(1+e^{-V(x)}\right) = -T(x)$
GAN-RKL	$-\log u$	$-V(x)$
GAN-KL	$u \log u$	$V(x)e^{V(x)}$
GAN-α	$\frac{1}{\alpha(\alpha-1)}\left(u^{\alpha}-1-\alpha(u-1)\right)$	$\frac{1}{\alpha(\alpha-1)}\left(e^{\alpha V(x)}-1-\alpha(e^{V(x)}-1)\right)$

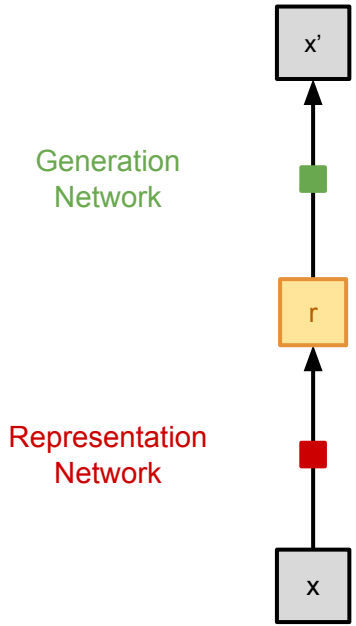
Mapping out the landscape of Generative Models

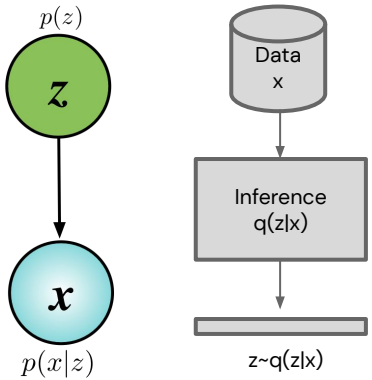


Spectrum of Latent Variable Models

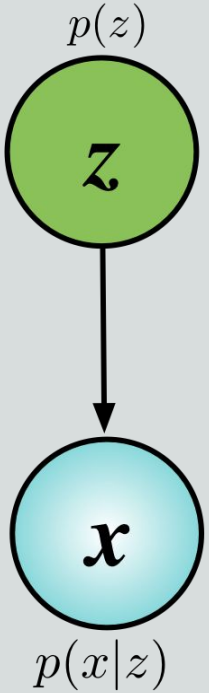


Representing Models: Computational graphs vs plate notation





Latent Variable Models

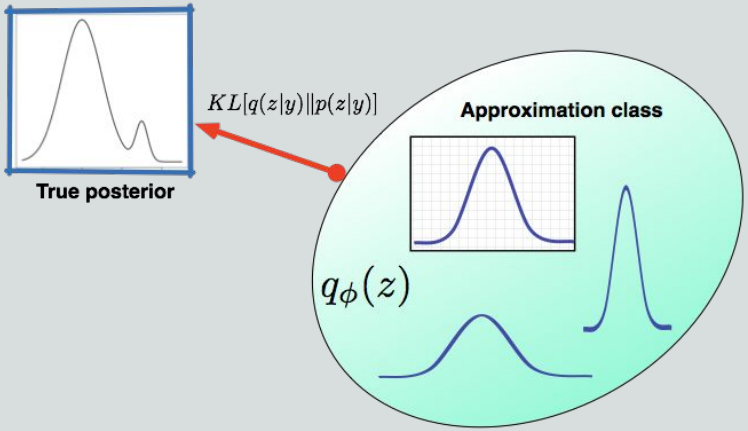


$$x \in \mathbb{R}^{d_x} \quad z \in \mathbb{R}^{d_z} \quad \theta \in \mathbb{R}^{d_\theta}$$
$$\mathcal{D} = \{x_i\} \quad i \in \{1, ..., N\}$$

$$\log p_\theta(x) = \log \int p_\theta(x|z)p(z)dz = \log \mathbb{E}_{p(z)}[p_\theta(x|z)]$$

$$\log p_\theta(\mathcal{D}) = \sum_{i=1}^{N} \log \mathbb{E}_{p(z)}[p_\theta(x_i|z)]$$

Variational Approximation



$$\log p_\theta(\mathcal{D}) = \sum_{i=1}^{N} \log \mathbb{E}_{p(z)}[p_\theta(x_i|z)]$$

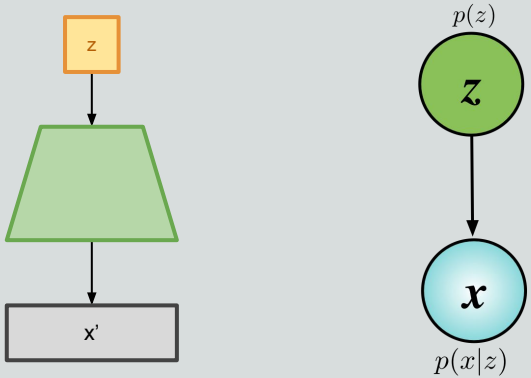
$$\log \mathbb{E}_{p(z)}[p_\theta(x_i|z)] = \log \mathbb{E}_{q_i(z)}\left[\frac{p_\theta(x_i|z)p(z)}{q_i(z)}\right], \quad \forall q_i > 0$$

$$\log \mathbb{E}_{q_i(z)}\left[\frac{p_\theta(x_i|z)p(z)}{q_i(z)}\right] \geqslant \mathbb{E}_{q_i(z)}\left[\log \frac{p_\theta(x_i|z)p(z)}{q_i(z)}\right]$$

$$\log p_\theta(\mathcal{D}) \geqslant \sum_{i=1}^{N} \mathbb{E}_{q_i(z)}\left[\log \frac{p_\theta(x_i|z)p(z)}{q_i(z)}\right]$$

Known as proposal, encoder or posterior model

Variational Inference: ELBO



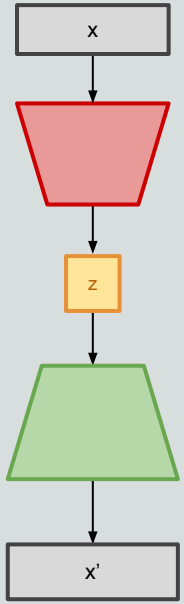
$$\log p_\theta(\mathcal{D}) \geqslant \sum_{i=1}^{N} \mathbb{E}_{q_i(z)}\left[\log \frac{p_\theta(x_i|z)p(z)}{q_i(z)}\right]$$

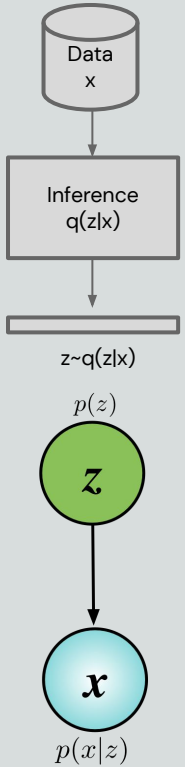
$$\mathbb{E}_{q_i(z)}\left[\log \frac{p_\theta(x_i|z)p(z)}{q_i(z)}\right] = \mathbb{E}_{q_i(z)}[\log p_\theta(x_i|z)] - \mathrm{KLD}(q_i\|p)$$

Reconstruction (distortion)

Regularizer (rate)

Amortised Inference





Want to learn more?

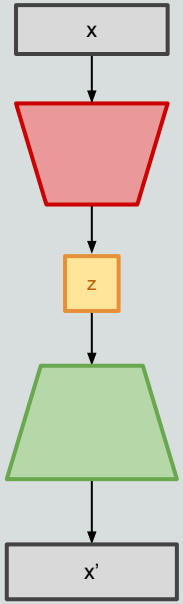
Stochastic Backpropagation and Approximate Inference in Deep Generative Models, Rezende et al, ICML 2014

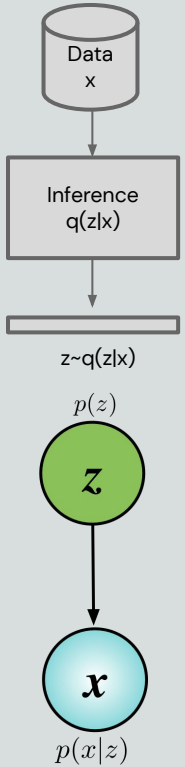
$$q_i^\star(z) = \text{argmax}_{q_i} \mathbb{E}_{q_i^\star(z)}[-\mathcal{F}(x_i, z)]$$

Introduce a parametric family of conditional densities

$$\text{argmax}_{q_i} \mathbb{E}_{q_i^\star(z)}[-\mathcal{F}(x_i, z)] \Rightarrow \text{argmax}_{\phi} \mathbb{E}_{q_\phi(z|x)}[-\mathcal{F}_\phi(x_i, z)$$

Variational AutoEncoder (VAE)





Want to learn more?

Stochastic Backpropagation and Approximate Inference in Deep Generative Models, Rezende et al, ICML 2014

Auto-Encoding Variational Bayes, Kingma & Welling, ICLR 2014

Deep Latent Gaussian Model p(x,z)

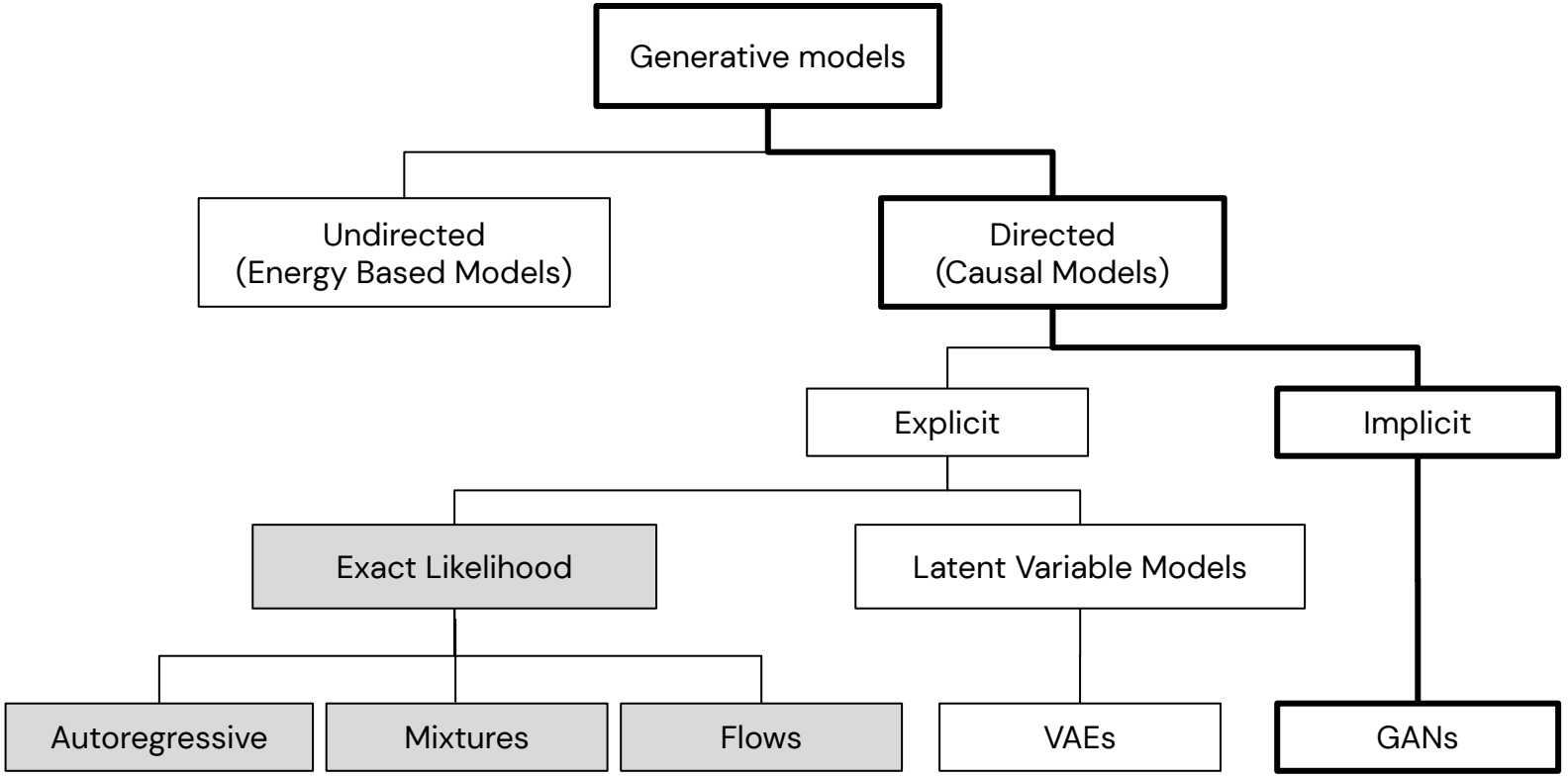
$$\begin{aligned} \text{prior sample} \quad & z \sim \mathcal{N}(0, \mathbb{I}) \\ \text{data sufficient statistics} \quad & \eta = f_\theta(z) \\ \text{data conditional likelihood} \quad & x \sim \mathcal{N}(\eta) \end{aligned}$$

Gaussian Recognition Model q(z)

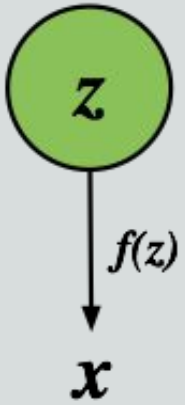
$$\begin{aligned} \text{data sample} \quad & x \sim \mathcal{D} \\ \text{latent sufficient statistics} \quad & \eta = f_\phi(x) \\ \text{posterior sample} \quad & z \sim \mathcal{N}(\eta) \end{aligned}$$

$$\mathbb{E}_{q_i(z)}[\log p_\theta(x_i|z)] - \mathrm{KLD}(q_i \| p)$$

Mapping out the landscape of Generative Models



Latent Variable Models



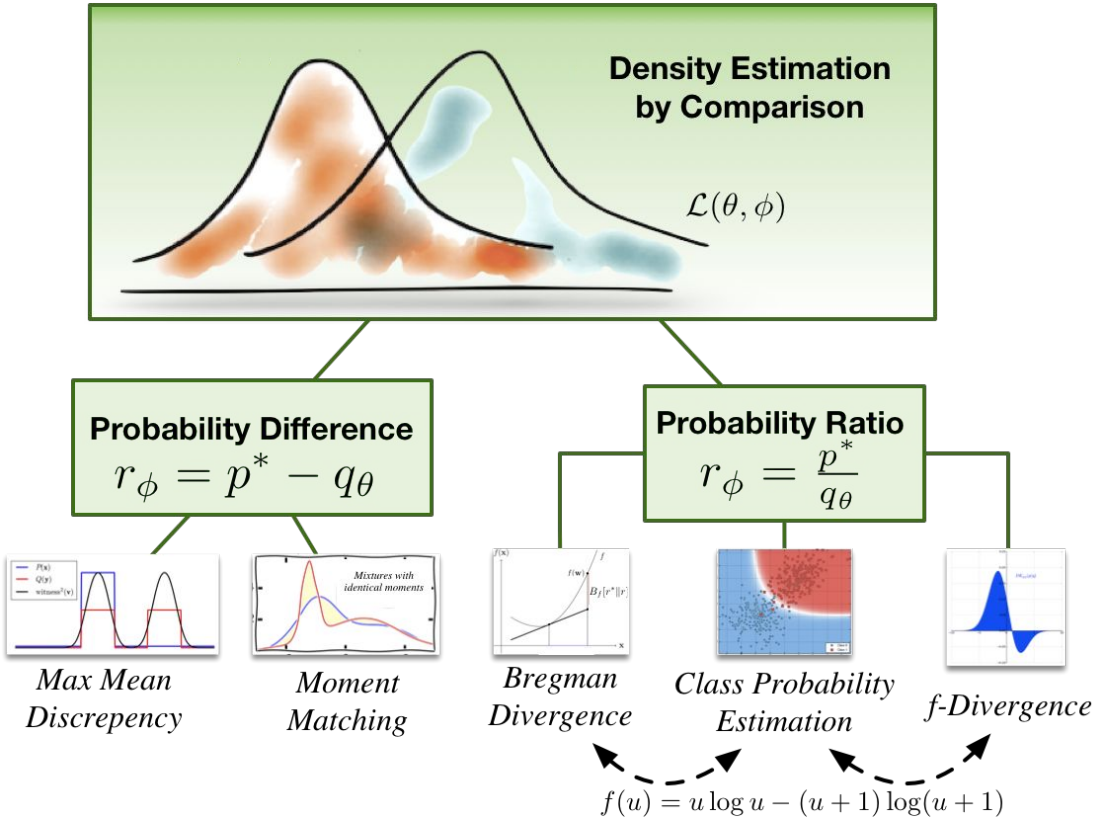
$$x \in \mathbb{R}^{d_x} \quad z \in \mathbb{R}^{d_z} \quad \theta \in \mathbb{R}^{d_\theta}$$
$$\mathcal{D} = \{x_i\} \quad i \in \{1, ..., N\}$$

$$\ln p_\theta(x) = \ln \int \delta(x - f(z)) p(z) dz$$

Learning by comparison

Want to learn more?

Learning in Implicit Generative Models, Mohamed and Lakshminarayanan, ICML 2017



Unsupervised-as-Supervised Learning

Want to learn more?

Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems. 2014.

Scoring Function

$$p(y=+1|\mathbf{x}) = D_\theta(\mathbf{x}) \qquad p(y=-1|\mathbf{x}) = 1 - D_\theta(\mathbf{x})$$

Bernoulli Loss

$$\mathcal{F}(\mathbf{x}, \theta, \phi) = \mathbb{E}_{p^*(x)}[\log D_\theta(\mathbf{x})] + \mathbb{E}_{q_\phi(x)}[\log(1 - D_\theta(\mathbf{x})]$$

Alternating optimisation

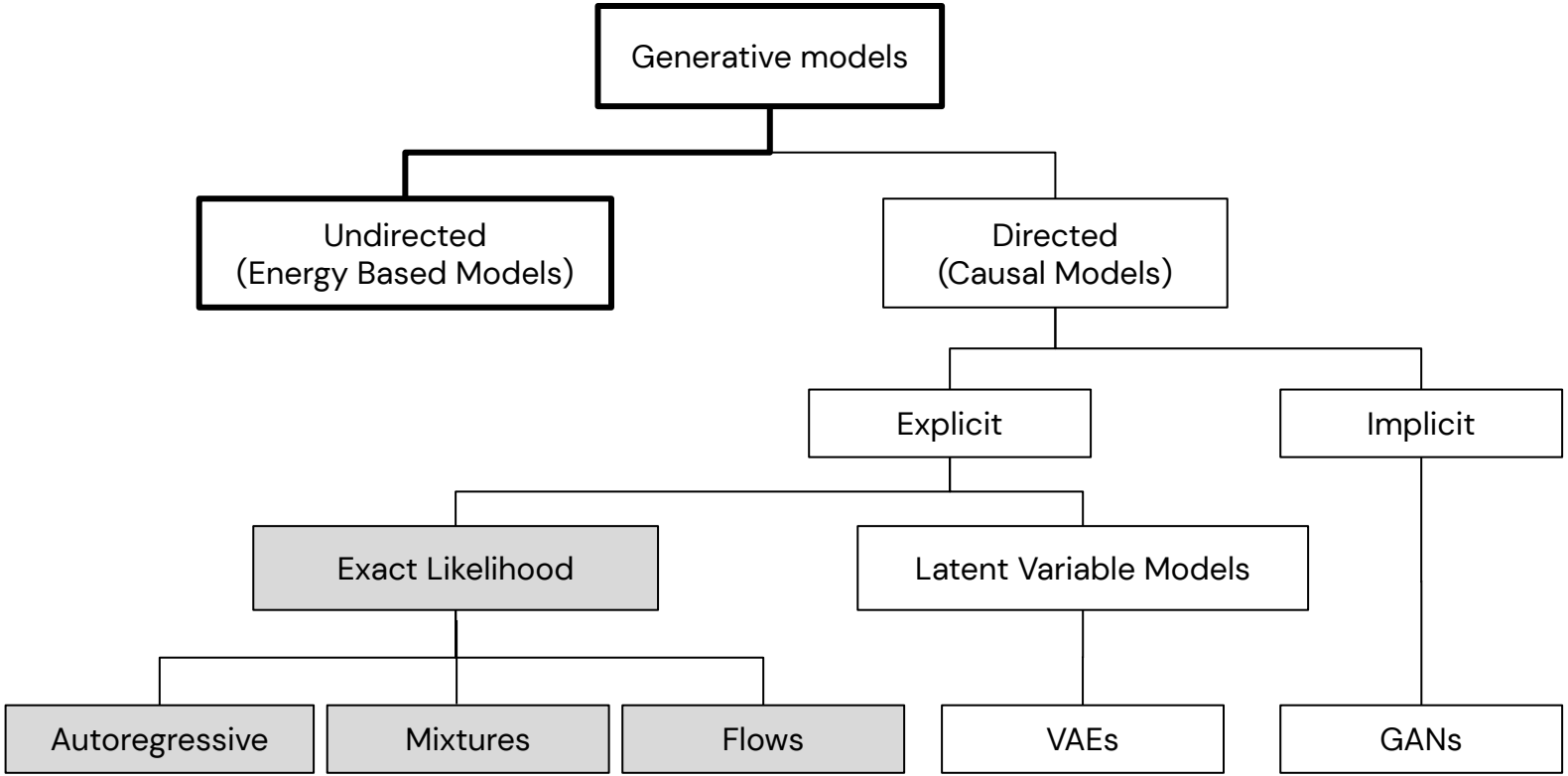
$$\min_\phi \max_\theta \mathcal{F}(\mathbf{x}, \theta, \phi)$$

- Use when we have differentiable simulators and models
- Can form the loss using any proper scoring rule.

Other names and places:

- Unsupervised and supervised learning
- Continuously updating inference
- Classifier ABC
- Generative Adversarial Networks

Mapping out the landscape of Generative Models

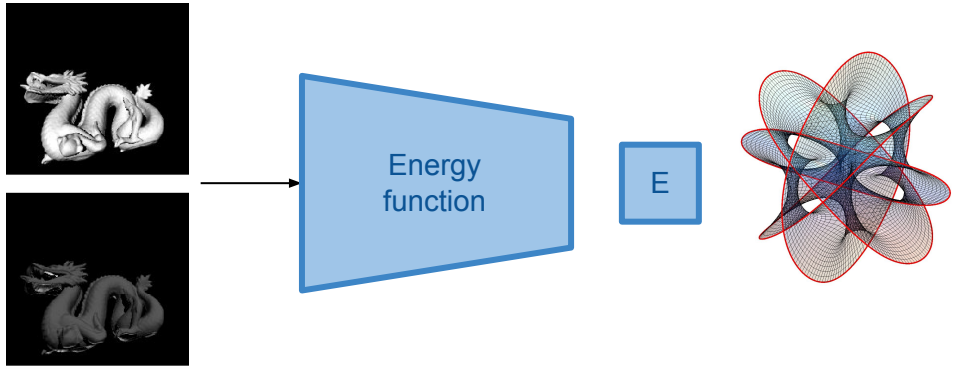


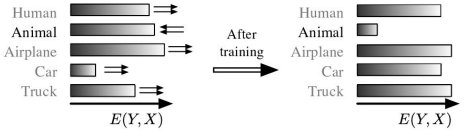
Energy models

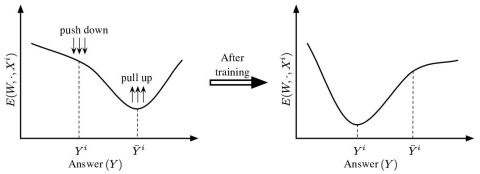
Want to learn more?

A Tutorial on Energy-Based Learning, LeCun et al, Predicting Structured Data 2006

- Learn energy manifold from data
- Choice of energy function (minimised during inference)
  - Implicit choice of metric
  - Shapes learnt manifold
- Choice of loss functional (minimised during learning)
  - Controls how hard energy manifold is shaped by contrastive examples





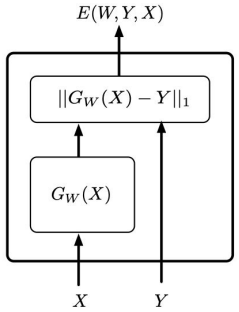


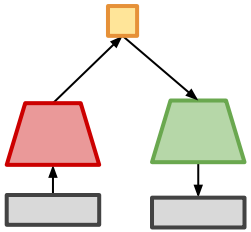
Energy models: $y = x^2$

Want to learn more?

A Tutorial on Energy-Based Learning, LeCun et al, Predicting Structured Data 2006

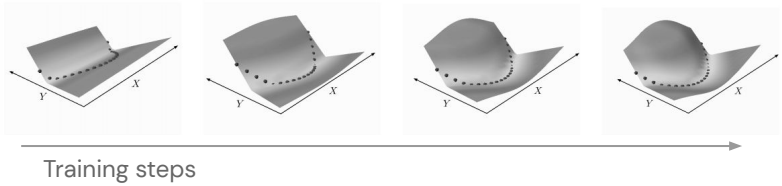
- Learn energy manifold from data
- Choice of energy function (minimised during inference)
  - Implicit choice of metric
  - Shapes learnt manifold
- Choice of loss functional (minimised during learning)
  - Controls how hard energy manifold is shaped by contrastive examples





Energy: $E(W,Y,X) = ||G_W(X) - Y||_1.$

Loss: $\mathcal{L}_{\text{energy}}(W,\mathcal{S}) = \frac{1}{P}\sum_{i=1}^{P} E(W,Y^i,X^i) = \frac{1}{P}\sum_{i=1}^{P} ||G_W(X) - Y||_1.$

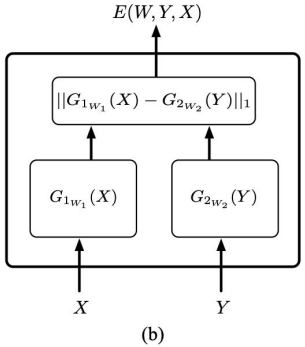


Energy models: $y = x^2$

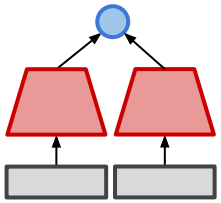
Want to learn more?

A Tutorial on Energy-Based Learning, LeCun et al, Predicting Structured Data 2006

- Learn energy manifold from data
- Choice of energy function (minimised during inference)
  - Implicit choice of metric
  - Shapes learnt manifold
- Choice of loss functional (minimised during learning)
  - Controls how hard energy manifold is shaped by contrastive examples

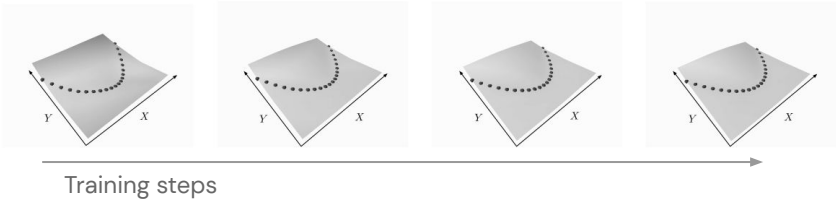


(b)



Energy: $E(W, X, Y) = ||G_{1_{W_1}}(X) - G_{2_{W_2}}(Y)||_1,$

Loss: $\mathcal{L}_{\text{energy}}(W, \mathcal{S}) = \frac{1}{P} \sum_{i=1}^{P} E(W, Y^i, X^i)$

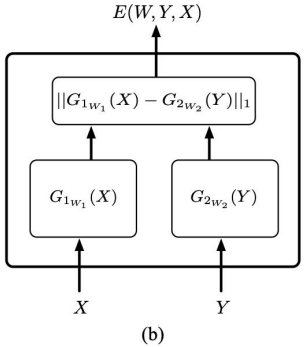


Energy models: $y = x^2$

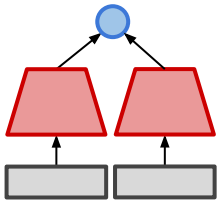
Want to learn more?

A Tutorial on Energy-Based Learning, LeCun et al, Predicting Structured Data 2006

- Learn energy manifold from data
- Choice of energy function (minimised during inference)
  - Implicit choice of metric
  - Shapes learnt manifold
- Choice of loss functional (minimised during learning)
  - Controls how hard energy manifold is shaped by contrastive examples

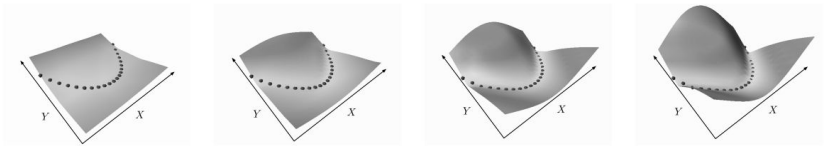


(b)



Energy: $E(W,X,Y) = ||G_{1_{W_1}}(X) - G_{2_{W_2}}(Y)||_1,$

Loss: $L(W,Y^i,X^i) = E(W,Y^i,X^i)^2 - \left(max(0, m - E(W,\bar{Y}^i,X^i))\right)^2.$

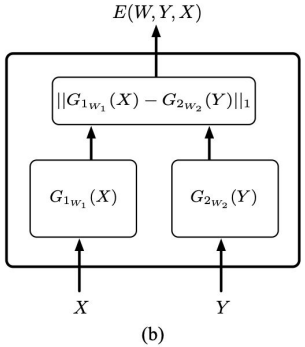


Energy models: $y = x^2$

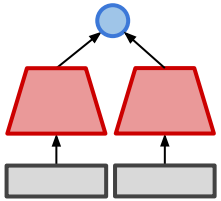
Want to learn more?

A Tutorial on Energy-Based Learning, LeCun et al, Predicting Structured Data 2006

- Learn energy manifold from data
- Choice of energy function (minimised during inference)
  - Implicit choice of metric
  - Shapes learnt manifold
- Choice of loss functional (minimised during learning)
  - Controls how hard energy manifold is shaped by contrastive examples

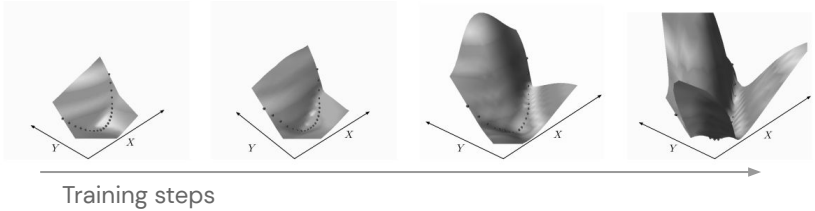


(b)



Energy: $E(W, X, Y) = ||G_{1_{W_1}}(X) - G_{2_{W_2}}(Y)||_1,$

Loss: $L(W, Y^i, X^i) = E(W, Y^i, X^i) + \frac{1}{\beta} \log \left(\int_{y \in \mathcal{Y}} e^{-\beta E(W, y, X^i)} \right).$



Maximum-Likelihood: Minimizing KL(data; model)

$$p(x;\theta) = Z(\theta)^{-1} e^{-E(x;\theta)}$$

$$Z(\theta) = \int dx e^{-E(x;\theta)}$$

$$\nabla_\theta \ln p(x;\theta) = -\nabla_\theta E(x;\theta) + \mathbb{E}_p[\nabla_\theta E(x';\theta)]$$

Easy:
evaluated at data

Hard:
evaluated at model samples

Sampling from energies: Langevin sampler

SDE: continuous time stochastic process

$$dX = -\nabla_x E(x;\theta)dt + \sqrt{2}d\xi$$

ξ is a stationary Gaussian process

$$\mathbb{E}[\xi(t)] = 0 \quad \mathbb{E}[\xi(t)\xi(t')] = \delta(t-t')$$

Basically, gradient descent with noise

$$x_{t+1} = x_t - \nabla_{x_t} E(x_t;\theta)dt + \sqrt{2dt}\xi$$

Will converge if **dt** → **0** and **t** → **∞**; but can easily get stuck in local minima

$$x_\infty \sim e^{-E(x)}$$

Langevin sampler: fixing discretisation errors

Want to learn more?

Optimal scaling of discrete approximations to Langevin diffusions, Roberts & Rosenthal, Journal of the Royal Statistical Society 2002

$$x_{t+1} = x_t - \nabla_{x_t} E(x_t;\theta)dt + \sqrt{2dt}\xi$$

$$x_{t+1} \sim q(x_{t+1}|x_t) = \mathbb{N}(x_{t+1}; x_t - \nabla_{x_t} E(x_t;\theta)dt, \sqrt{2dt})$$

Accept samples with probability

$$\alpha := \min\left\{1, \frac{q(x_t|x_{t+1})}{q(x_{t+1}|x_t)} e^{-E(x_{t+1})+E(x_t)}\right\}$$

Optimal dt must be chosen such that **E[α]=0.574**

Sampling from energies: Hamiltonian Monte-Carlo sampler

$$H(x,p) = E(x) + K(p)$$

$$\frac{dx}{dt} = \frac{\partial H(x,p)}{\partial p} = \frac{\partial K(p)}{\partial p}$$

$$\frac{dp}{dt} = -\frac{\partial H(x,p)}{\partial x} = -\frac{\partial E(x)}{\partial x}$$

$$p(t=0) \sim \mathbb{N}(0, \mathbb{I})$$

$$(x_\infty, p_\infty) \sim e^{-H(x,p)} = e^{-E(x)-K(p)}$$

$$\Rightarrow x_\infty \sim e^{-E(x)}$$

Sampling from energies: Hamiltonian Monte-Carlo sampler

Want to learn more?

MCMC using Hamiltonian Dynamics, Neal, Handbook of Markov Chain Monte Carlo 2011

Use leap-frog

$$p(t+\frac{dt}{2}) = p(t)\frac{dt}{2}\nabla_x E(x)$$

$$x(t+dt) = x(t) + dt\nabla_{p(t+\frac{dt}{2})} K(p(t+\frac{dt}{2}))$$

$$p(t+dt) = p(t+\frac{dt}{2}) - \frac{dt}{2}\nabla_{x(t+dt)} E(x(t+dt))$$

Accept with probability

$$\alpha = \min\left\{1, e^{H(x,p)-H(x',p')}\right\}$$

Score Matching:
Minimizing the Fisher divergence

$$q(x;\theta) = Z^{-1} e^{-E(x;\theta)}$$

$$Z(\theta) = \int dx e^{-E(x;\theta)}$$

Usually unknown

$$\begin{aligned}\mathrm{FD}(p,q) &= \mathbb{E}_p[\|\nabla_x \ln p(x) - \nabla_x \ln q(x;\theta)\|^2] \\ &= 2\mathbb{E}_p[\mathrm{Tr}[\nabla_x^2 \ln q(x;\theta)] + \|\nabla_x \ln q(x;\theta)\|^2] + \mathrm{cst}\end{aligned}$$

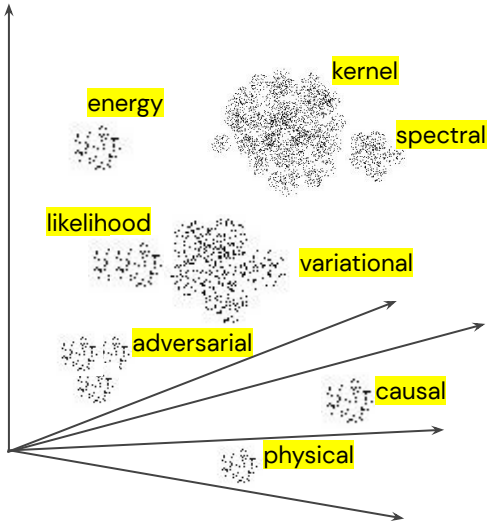
Integration by parts, **p**, **q** and their gradients must go to zero at the boundary

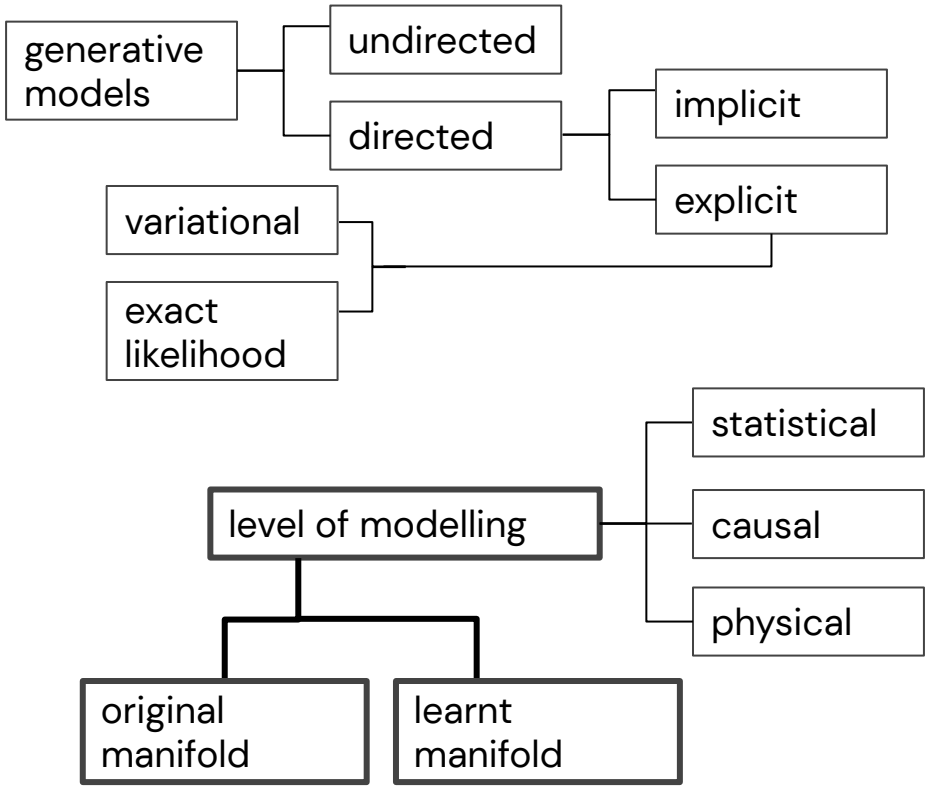
Easy: average over data

Expensive: Hessian

Good: Does not depend on normalizer **z**

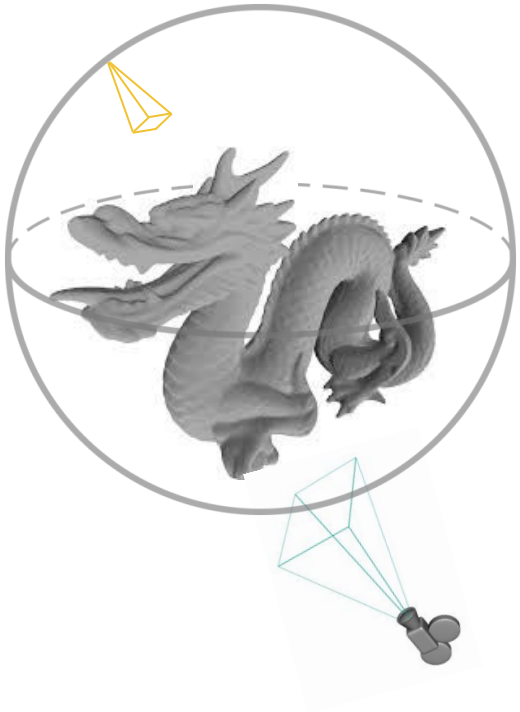
Mapping out the landscape



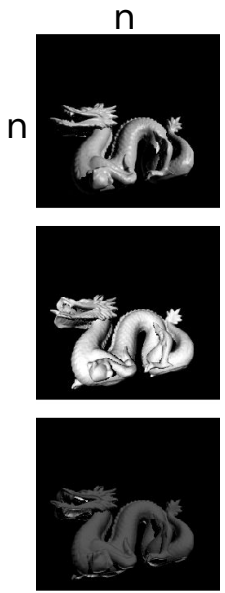


Manifold hypothesis

"Real-world high dimensional data lie on low-dimensional manifolds embedded in the high-dimensional space."

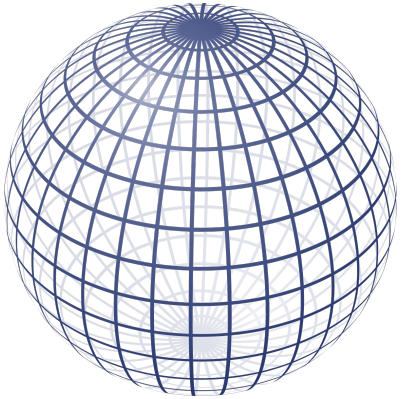


Pixel space



$$\mathbb{R}^{n^2}$$

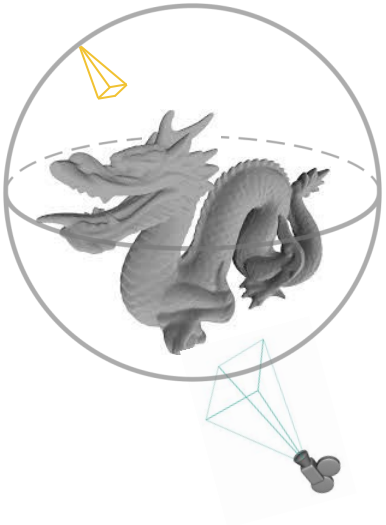
Sphere



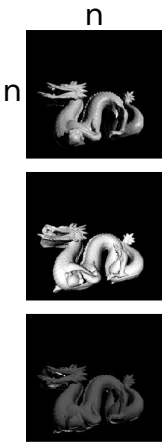
$$\mathrm{S}^2 \in \mathbb{R}^3$$

Manifold hypothesis

"Real-world high dimensional data lie on low-dimensional manifolds embedded in the high-dimensional space."

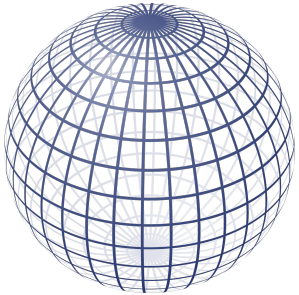


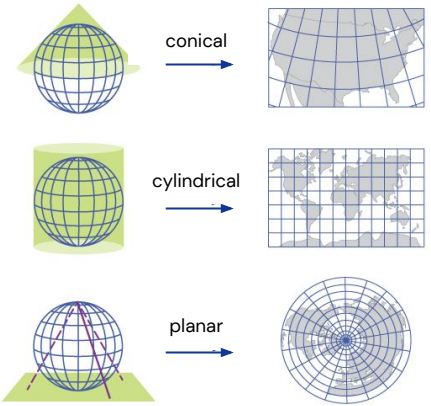
Pixel space



$\mathbb{R}^{n^2}$

Sphere





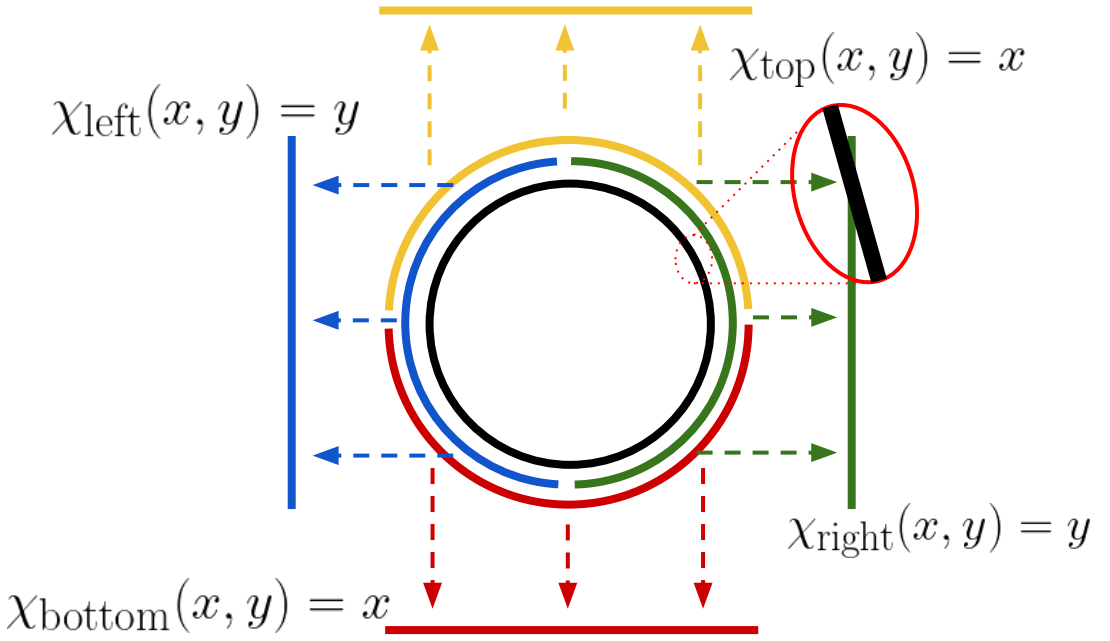
$\mathrm{S}^2 \in \mathbb{R}^3 \longrightarrow \mathbb{R}^2$

Manifolds

Want to learn more?

Manifolds: A Gentle Introduction, Keng, 2018
http://bjlkeng.github.io/posts/manifolds/

- Topological space that "locally" resembles Euclidean space.
- Functions and open regions they map are called ***charts***
- Set of all charts that map the whole manifold make an ***atlas***



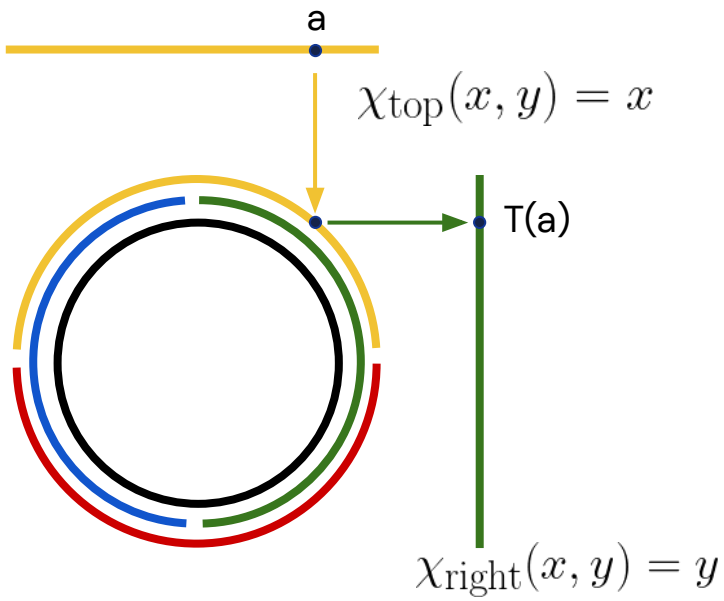
Manifolds

Want to learn more?

Manifolds: A Gentle Introduction, Keng, 2018
http://bjlkeng.github.io/posts/manifolds/

- Topological space that "locally" resembles Euclidean space.
- Functions and open regions they map are called ***charts***
- Set of all charts that map the whole manifold make an ***atlas***
- To move between charts, use ***transition maps***

$$T : (0,1) \to (0,1) = \chi_{\text{right}} \circ \chi_{\text{top}}^{-1}$$



Manifolds

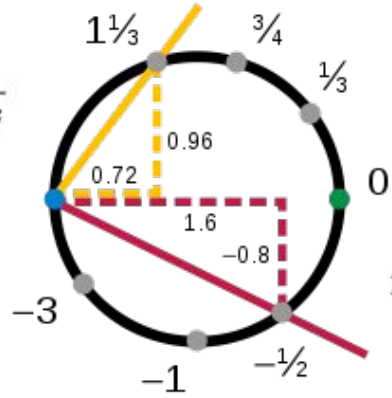
Want to learn more?

Manifolds: A Gentle Introduction, Keng, 2018
http://bjlkeng.github.io/posts/manifolds/

- Topological space that "locally" resembles Euclidean space.
- Continuous and invertible functions from segment to open region are called ***charts***
- Set of all charts that cover the whole manifold make an ***atlas***
- To move between charts, use ***transition maps***
- Many possible choices for charts and atlases

$$\chi_{\text{plus}}(x, y) = t = \frac{y}{1 - x}$$

$$t = \frac{1}{s}$$



$$\chi_{\text{minus}}(x, y) = s = \frac{y}{1 + x}$$

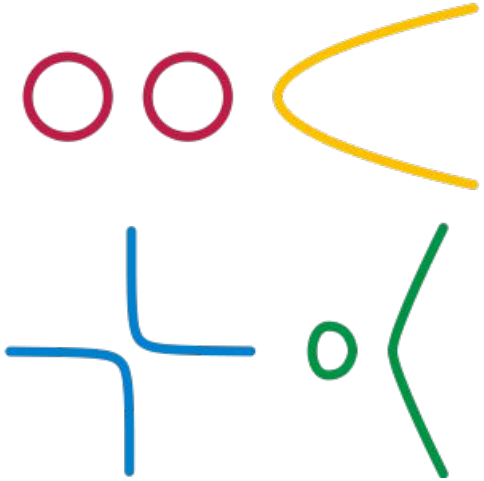
Manifolds

Want to learn more?

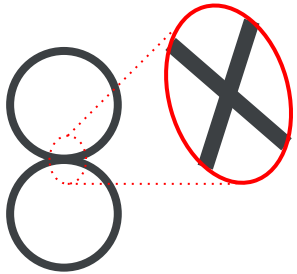
Manifolds: A Gentle Introduction, Keng, 2018
http://bjlkeng.github.io/posts/manifolds/

- Topological space that "locally" resembles Euclidean space.
- Homeomorphic maps from subset of manifold to Euclidean space R^n are called ***charts***
- Set of all charts that cover the whole manifold make an ***atlas***
- To move between charts, use ***transition maps***
- Many possible choices for charts and atlases
- Manifolds need not be ***connected*** or ***closed***

1D Manifolds



Not a manifold



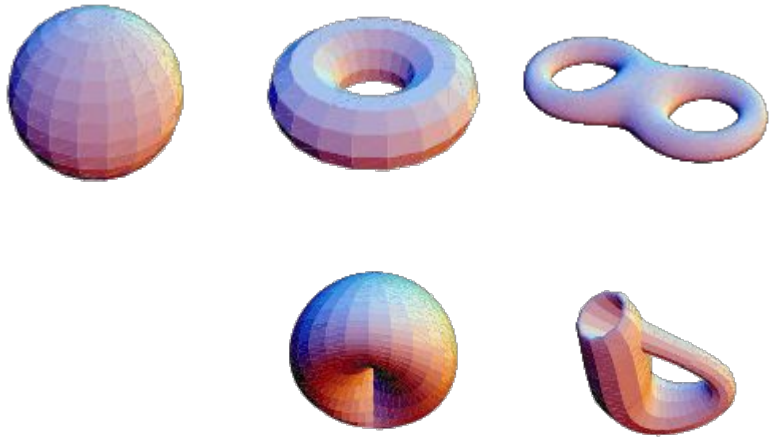
Manifolds

Want to learn more?

Manifolds: A Gentle Introduction, Keng, 2018
http://bjlkeng.github.io/posts/manifolds/

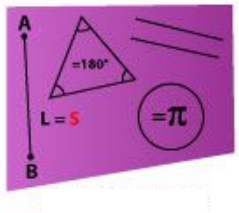
- Topological space that "locally" resembles Euclidean space.
- Homeomorphic maps from subset of manifold to Euclidean space R^n are called ***charts***
- Set of all charts that cover the whole manifold make an ***atlas***
- To move between charts, use ***transition maps***
- Many possible choices for charts and atlases
- Manifolds need not be ***connected*** or ***closed***

2D Manifolds

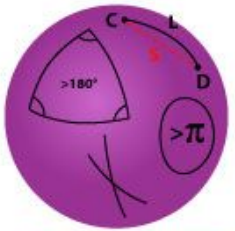


Euclidean space as a manifold

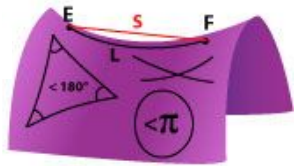
- R^n is a manifold
- Single chart = identity function
- Atlas contains single chart
- Distance = inner product



No curvature



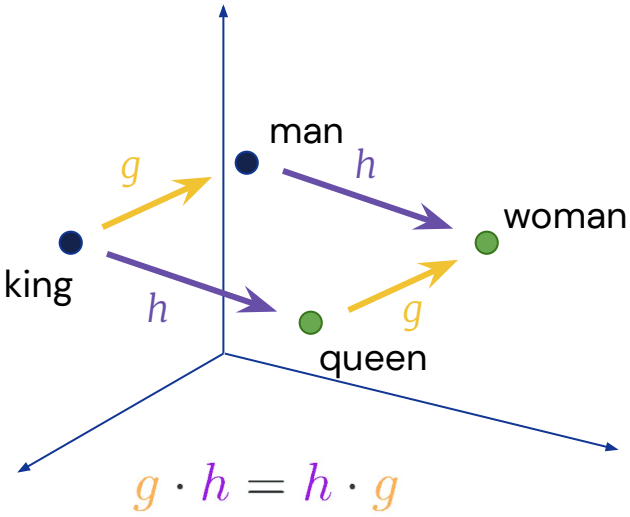
Positive curvature



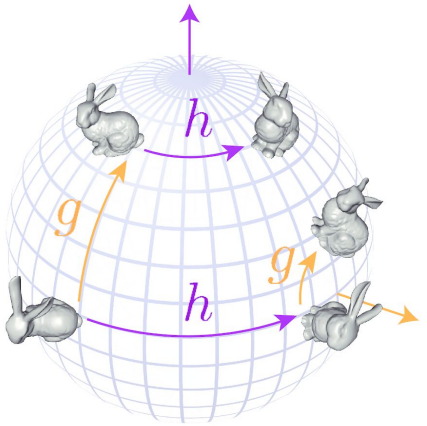
Negative curvature

Euclidean space as a manifold

- R^n is a manifold
- Single chart = identity function
- Atlas contains single chart
- Distance = inner product



$$g \cdot h = h \cdot g$$



$$g \cdot h \neq h \cdot g$$

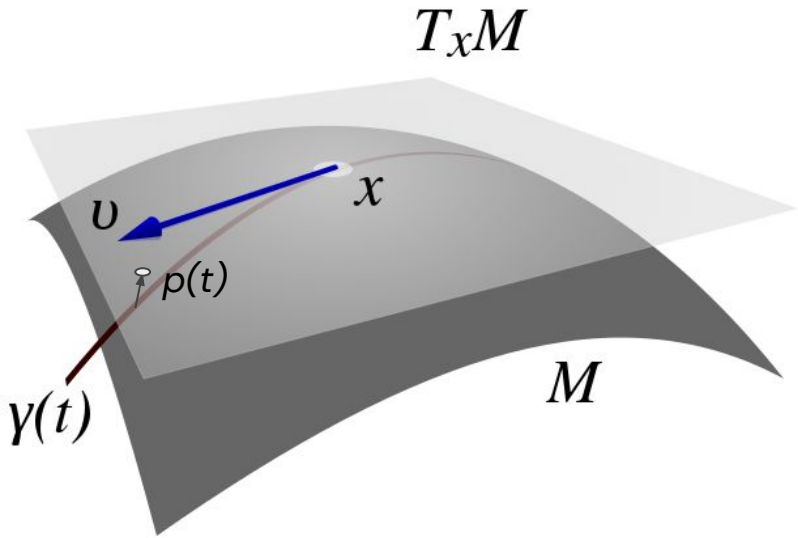
Want to learn more?

Efficient Estimation of Word Representations in Vector Space, Mikolov et al, ICLR 2013

Disentangling by Subspace Diffusion, Pfau et al, arxiv 2020

Tangent space

- Chart φ: **U** → $\mathbf{R}^n$ where **U** is an open subset of **M** containing **x**
- Curve γ: **t** → **M** runs along manifold through point **x**
- $\mathbf{p} = \varphi \circ \gamma(\mathbf{t})$ is position vector in $\mathbf{R}^n$
- $\mathbf{v} = d\mathbf{p}/d\mathbf{t} = (d\varphi \circ \gamma(\mathbf{t}))/d\mathbf{t}$, where $\mathbf{t} = \mathbf{t}_0$ is the "velocity" at **x**
- $\mathbf{v} \in \mathbf{R}^n$ is the tangent vector at **x**
- Tangent space $\mathbf{T}_x\mathbf{M}$ on manifold **M** at point **x** is the collection of all tangent vectors $\mathbf{v} \in \mathbf{R}^n$



Riemannian metric

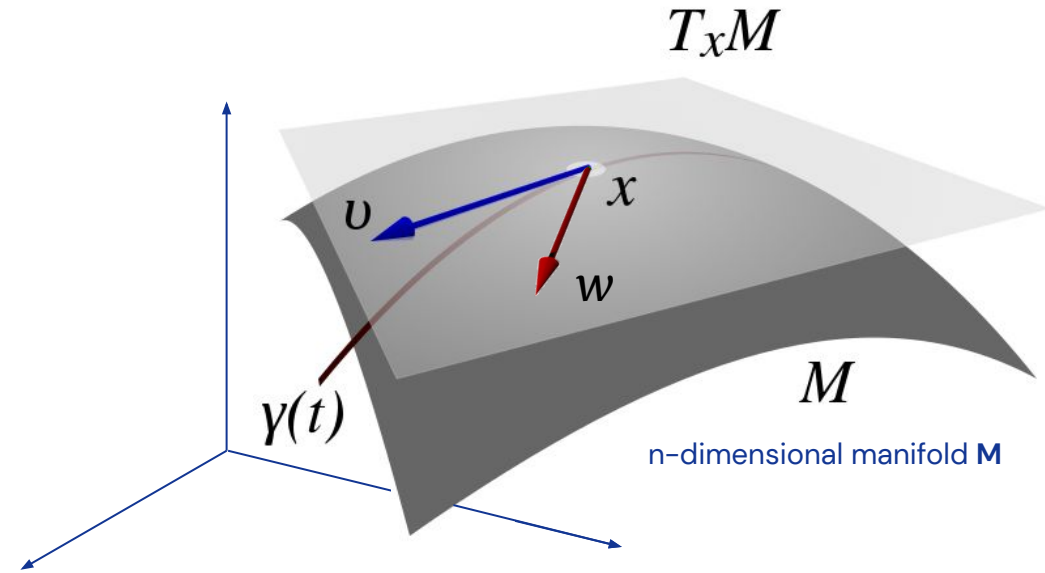
- Necessary to calculate distances (geodesics), angles, area etc.
- Riemannian metric is a family of function:

  $g_x : T_xM \times T_xM \rightarrow R, x \in M$

  such that $x \rightarrow g_x(v, w)$ is a smooth function of **x.**
- Generalisation of inner product:

  $g_x(v, w) = v^T J_q{}^T J_q w$

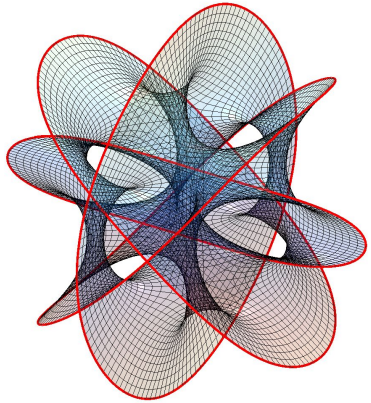
  where J_q is Jacobian of function $q: T_xM \rightarrow R^{n+k}$
- Different metric for every point on the manifold - tensor field



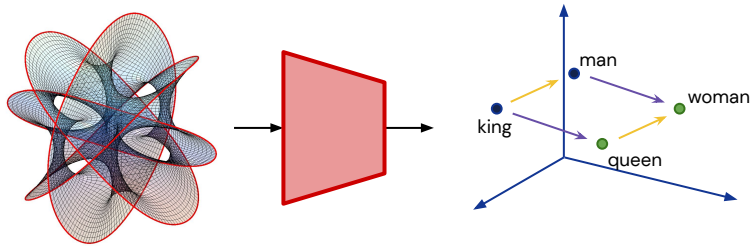
n-dimensional manifold **M**

embedded in (n+k)-dimensional Euclidean space

Two options for building models

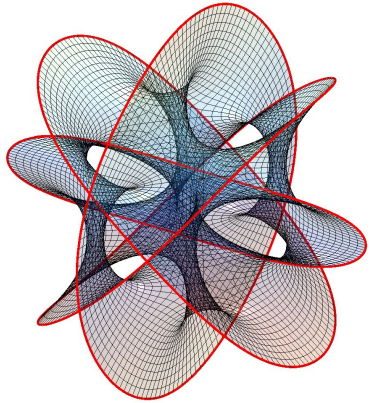


- Stay in original data manifold
- Learn appropriate atlas and metric

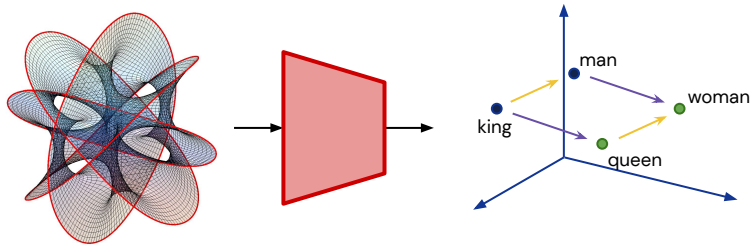


- Move away from original data manifold
- Assume (Euclidean) manifold with simple atlas and metric
- Learn projection to such manifold

Two options for building models



- Stay in original data manifold
- Learn appropriate atlas and metric



- Move away from original data manifold
- Assume (Euclidean) manifold with simple atlas and metric
- Learn projection to such manifold

Spectral/kernel methods

Want to learn more?

Nonlinear Component Analysis as a Kernel Eigenvalue Problem, Schölkopf, Neural Computation 1998

A Global Geometric Framework for Nonlinear Dimensionality Reduction, Tennenbaum et al, Science 2000

Visualizing Data using t-SNE, van der Maaten & Hinton, JMLR 2008

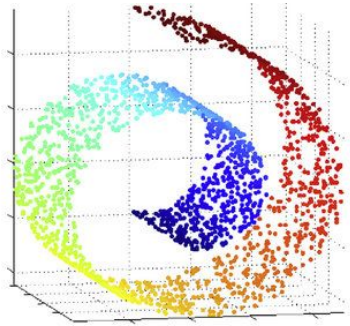
- Calculate data similarity
  - Covariance (PCA)
  - Euclidean distance (MDS, Isomap, LLE, Laplacian Eigenmap)
  - Stochastic neighbor embedding (t-SNE, UMAP)

- Perform spectral decomposition (PCA, kernel PCA, MDS, Isomap, LLE, Laplacian Eigenmap)

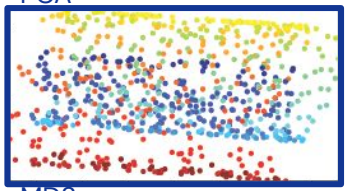
  and/or

- Optimise soft loss (MDS, Isomap, t-SNE, UMAP)

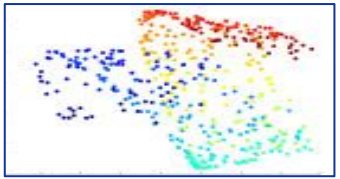
Original data



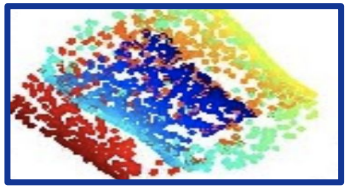
PCA



kernel PCA

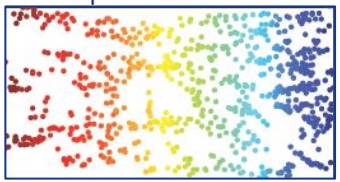


MDS

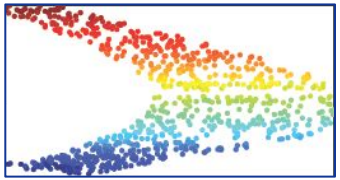


t-SNE

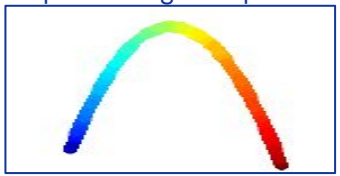
Isomap



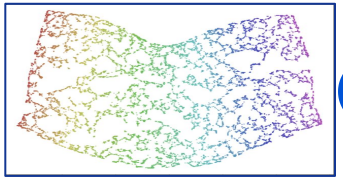
LLE



Laplacian Eigenmap



UMAP



Spectral/kernel methods

Want to learn more?

Nonlinear Component Analysis as a Kernel Eigenvalue Problem, Schölkopf, Neural Computation 1998

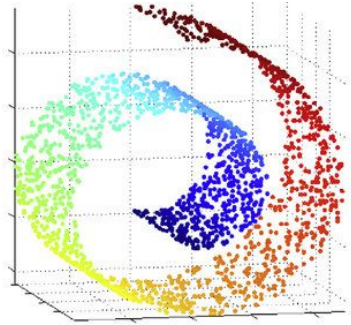
A Global Geometric Framework for Nonlinear Dimensionality Reduction, Tennenbaum et al, Science 2000

Visualizing Data using t-SNE, van der Maaten & Hinton, JMLR 2008

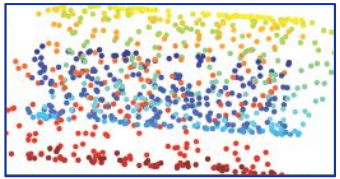
- Calculate data similarity
  - Covariance (PCA)
  - Euclidean distance (MDS, Isomap, LLE, Laplacian Eigenmap)
  - Stochastic neighbor embedding (t-SNE, UMAP)
  - Kernels (kernel PCA)
- (optional) Build neighbour graph (Isomap, LLE, t-SNE, UMAP, Laplacian Eigenmap)
- Perform spectral decomposition (PCA, kernel PCA, MDS, Isomap, LLE, Laplacian Eigenmap)

  and/or
- Optimise soft loss (MDS, Isomap, t-SNE, UMAP)

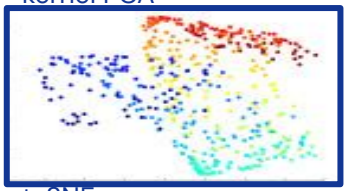
Original data



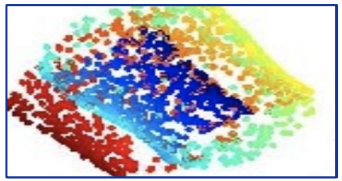
PCA



kernel PCA

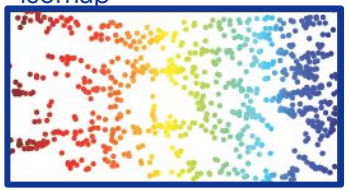


MDS



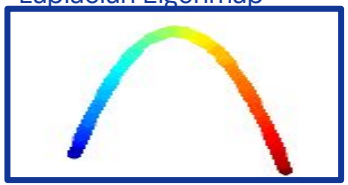
t-SNE

Isomap

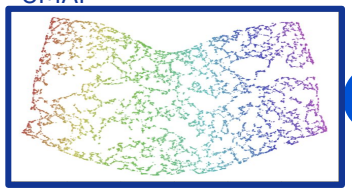


LLE

Laplacian Eigenmap



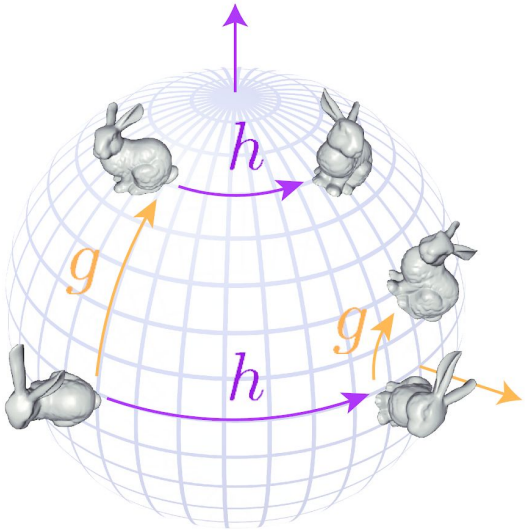
UMAP



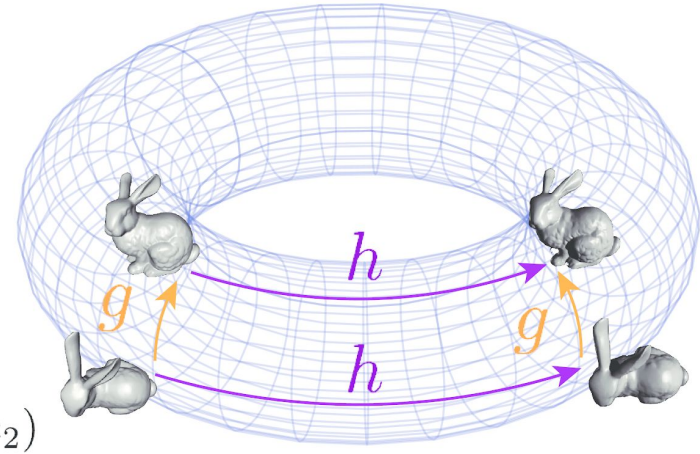
Geometric Manifold Component Estimator (GEOMANCER)

Want to learn more?

Disentangling by Subspace Diffusion, Pfau et al, arXiv 2020



$$g \cdot h \neq h \cdot g$$



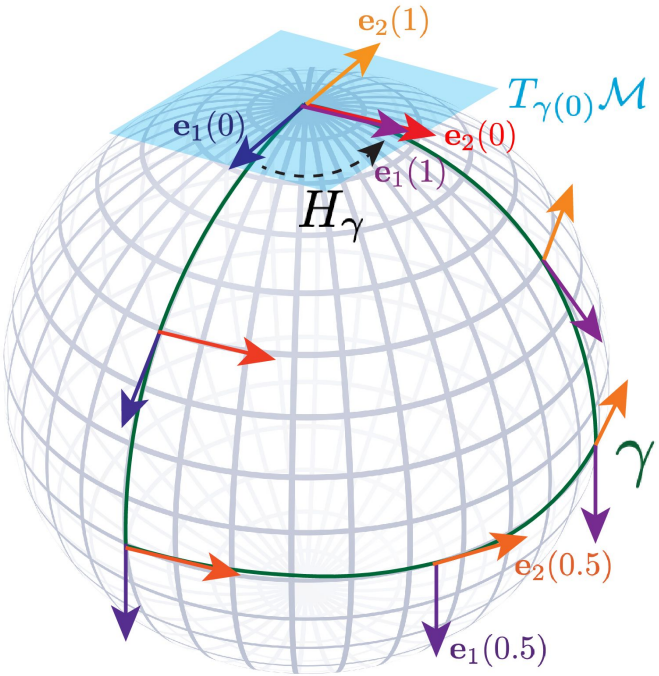
$$g = (g_1, e_2)$$
$$h = (e_1, h_2)$$

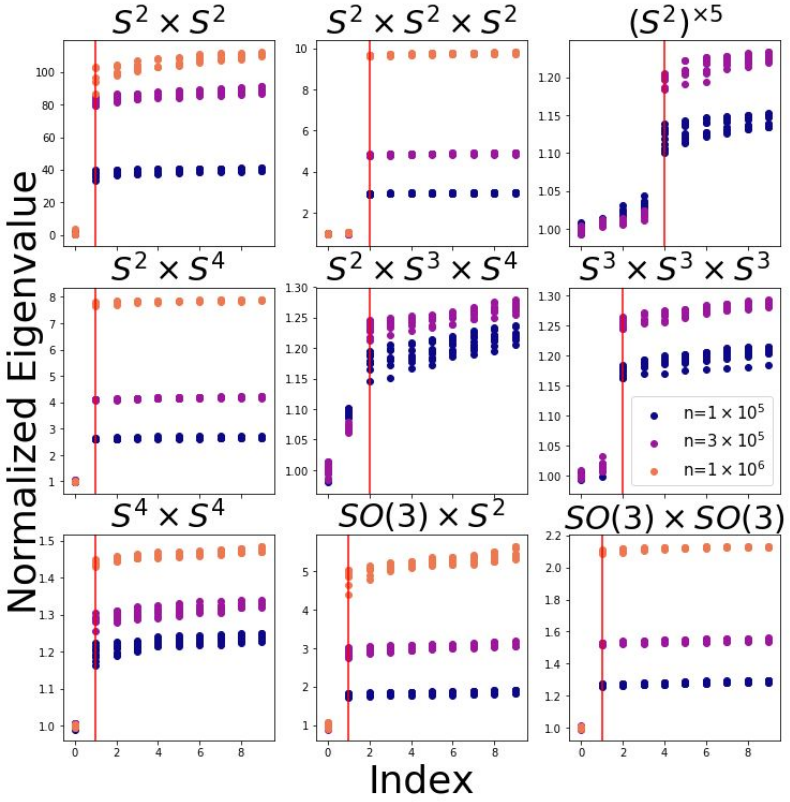
$$g \cdot h = h \cdot g = (g_1, h_2)$$

Geometric Manifold Component Estimator (GEOMANCER)

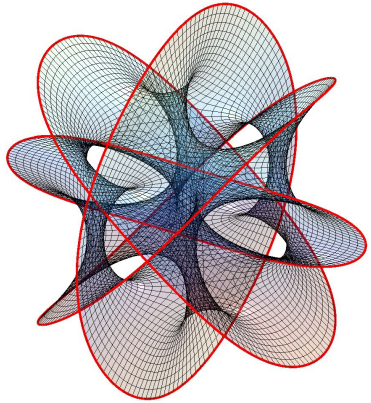
Want to learn more?

Disentangling by Subspace Diffusion, Pfau et al, arXiv 2020

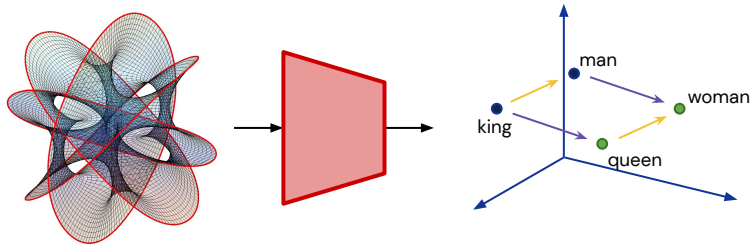




Two options for building models



- Stay in original data manifold
- Learn appropriate atlas and metric



- Move away from original data manifold
- Assume (Euclidean) manifold with simple atlas and metric
- Learn projection to such manifold

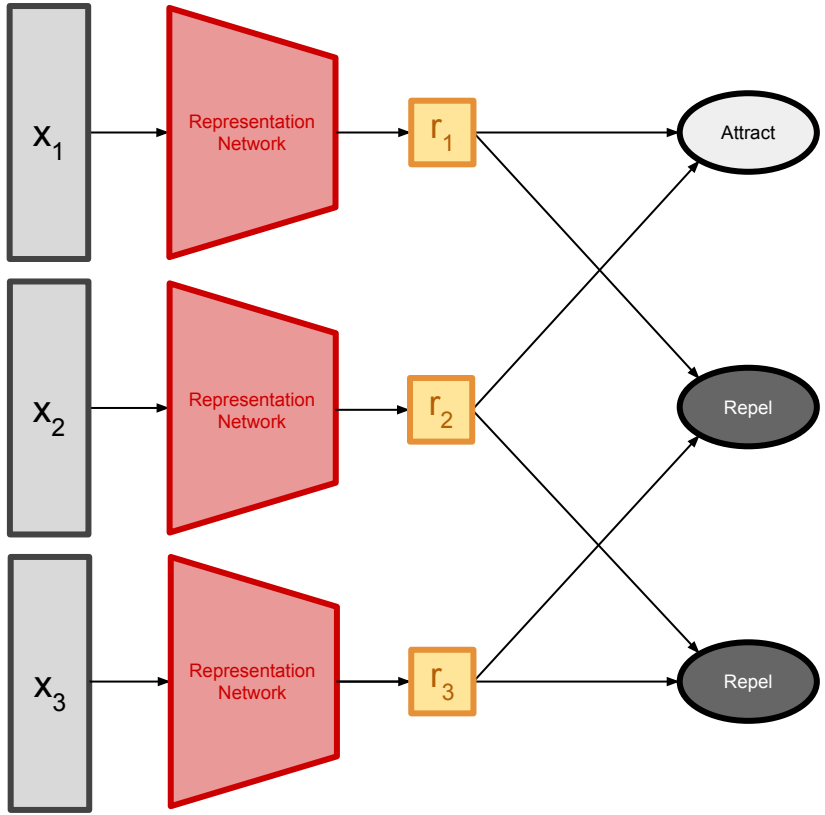
Contrastive learning / energy models / metric learning

Want to learn more?

Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning, Grill et al, arxiv 2020

Data-Efficient Image Recognition with Contrastive Predictive Coding, Hénaff et al, ICML 2020

A Simple Framework for Contrastive Learning of Visual Representations, Chen et al, ICML 2020



Positive pair

Negative pair

Negative pair

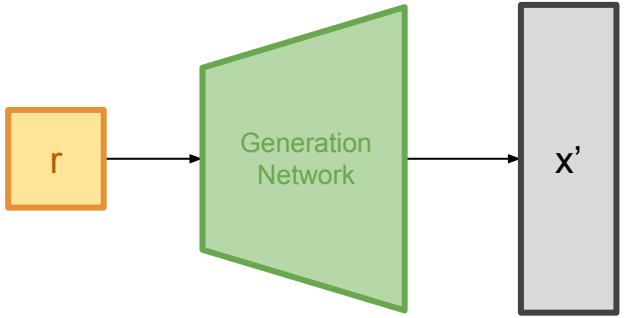
$$\mathcal{L}_{\text{SimCLR}} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)},$$

$$\text{sim}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v} / \|\boldsymbol{u}\| \|\boldsymbol{v}\|$$

$$\mathcal{L}_\theta^{\texttt{BYOL}} \triangleq \left\| \overline{q_\theta}(z_\theta) - \overline{z}'_\xi \right\|_2^2$$

$$\mathcal{L}_{\text{CPC}} = -\sum_{i,j,k} \log \frac{\exp(\hat{\boldsymbol{z}}_{i+k,j}^T \boldsymbol{z}_{i+k,j})}{\exp(\hat{\boldsymbol{z}}_{i+k,j}^T \boldsymbol{z}_{i+k,j}) + \sum_l \exp(\hat{\boldsymbol{z}}_{i+k,j}^T \boldsymbol{z}_l)}$$

Adversarial learning

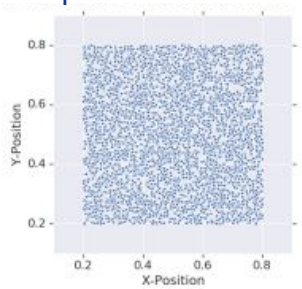


Variational Autoencoder learning

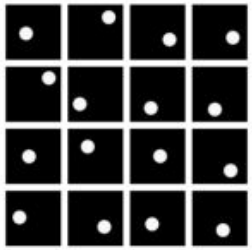
Want to learn more?

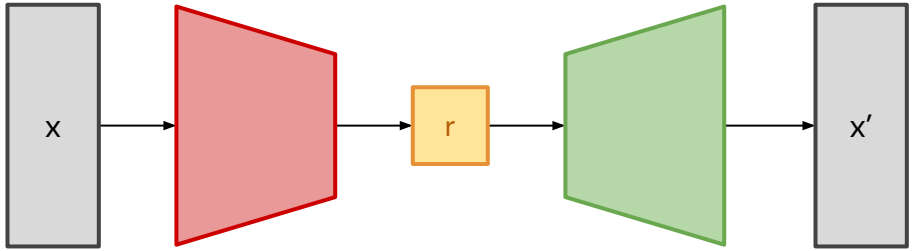
Spatial Broadcast Decoder: A Simple Architecture for Learning Disentangled Representations in VAEs, Watters et al, ICLR workshop 2018

Training data points

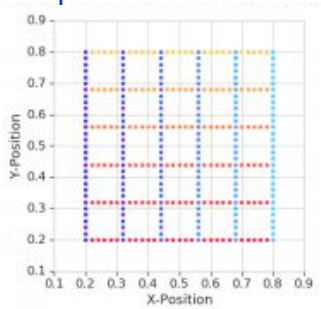


Dataset samples



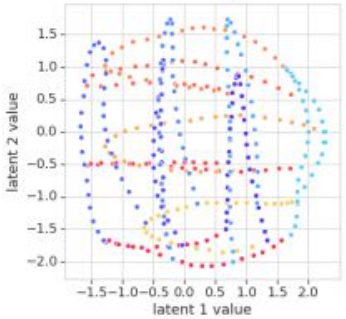


Test data points

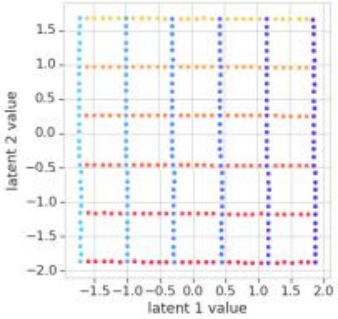


Pretrained encoder

VAE



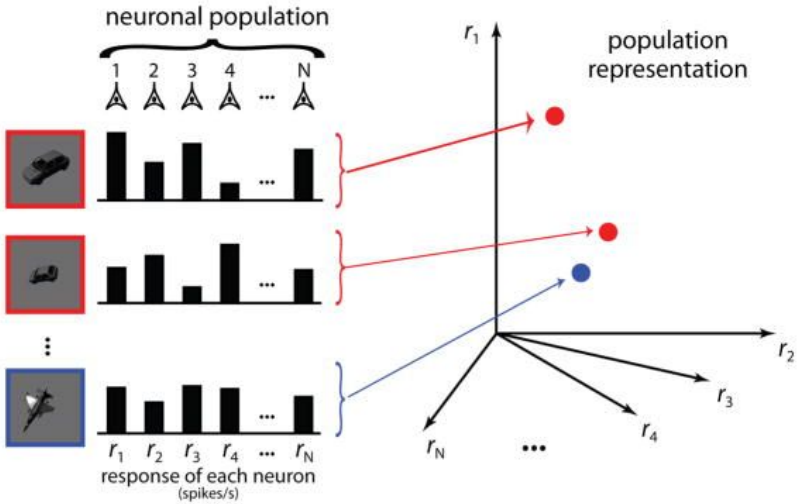
Factor VAE w\ BD



Untangling representations

Want to learn more?

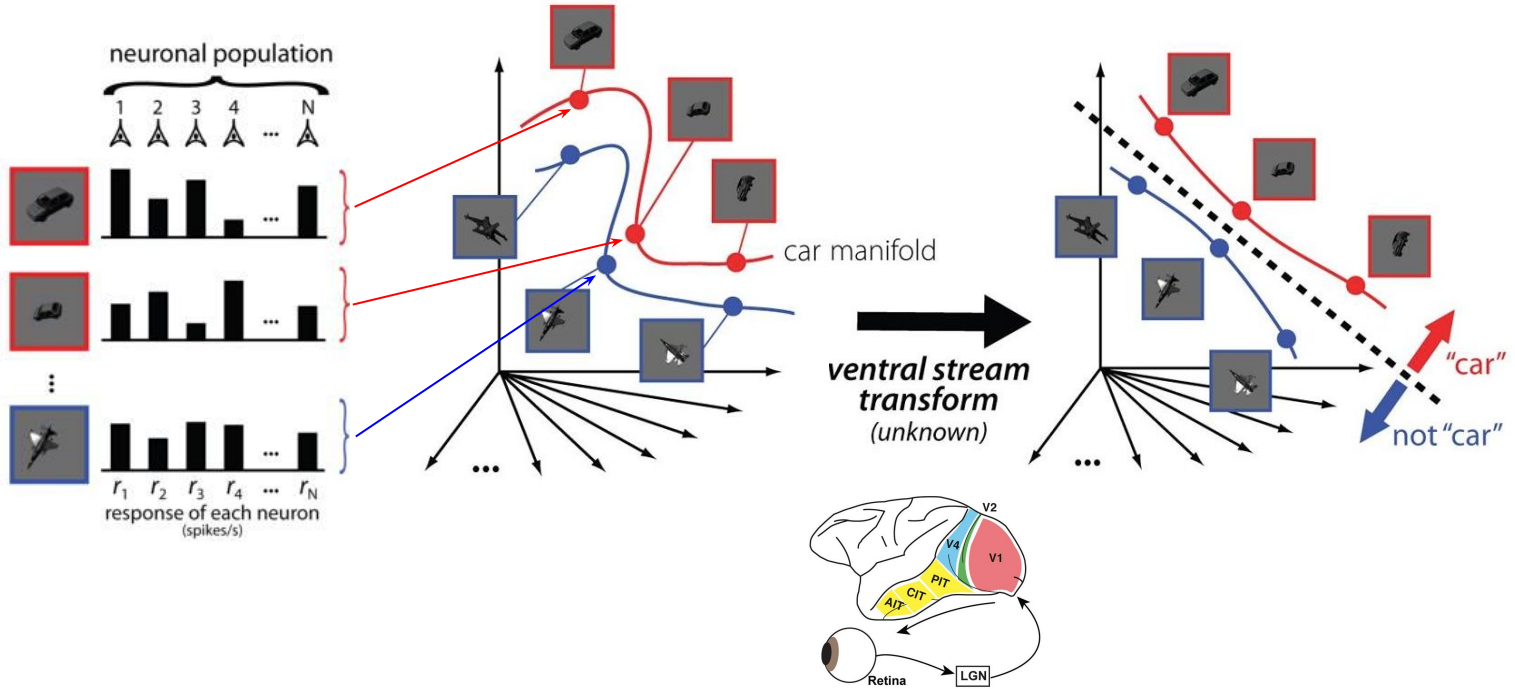
How Does the Brain Solve Visual Object Recognition?, DiCarlo et al, Neuron 2012



Untangling representations

Want to learn more?

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



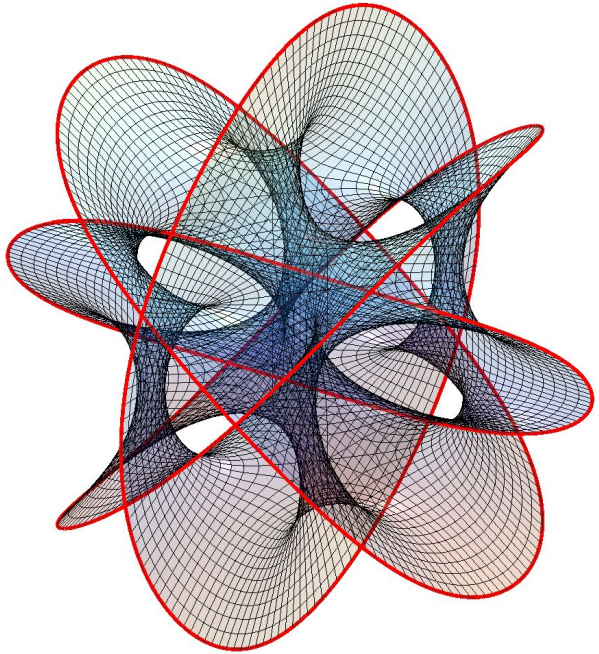
Summary

Thinking about the topology of the original **data manifold**,

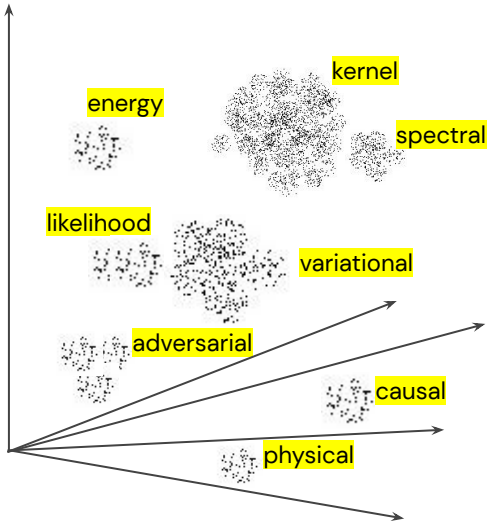
and the explicit or implicit **restrictions** on the topology of the learnt **representation manifold**

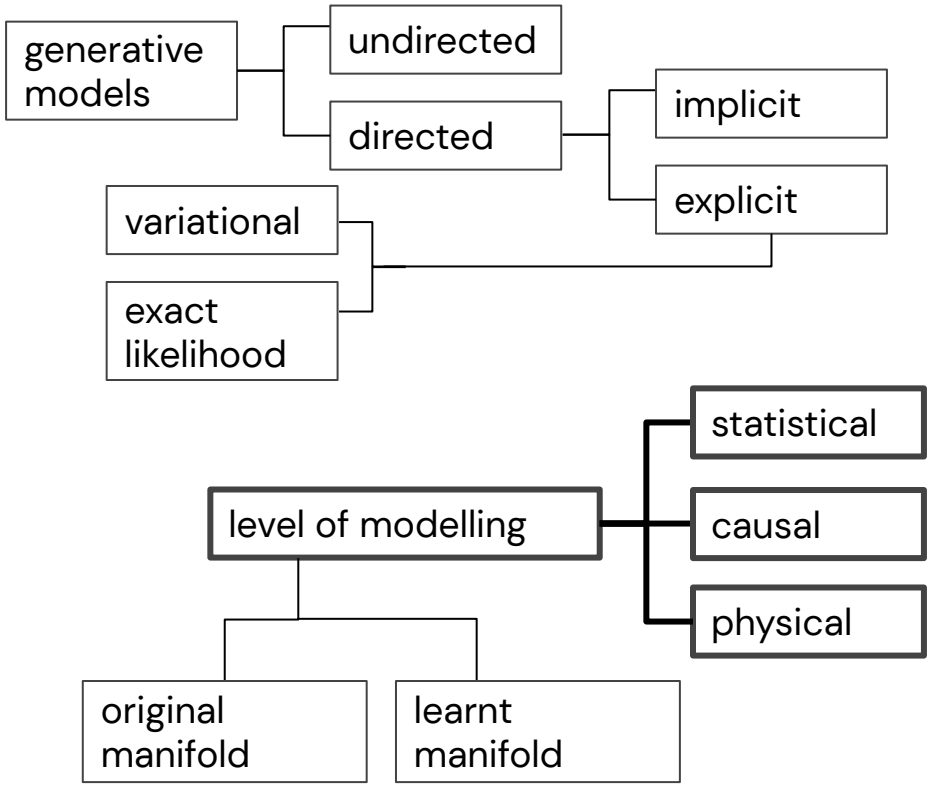
may help **troubleshoot problems** with the existing representation learning approaches

and help **develop better methods**.



Mapping out the landscape





Levels of modelling

Want to learn more?

Causality for Machine Learning, Schölkopf, arxiv 2019

$$p(\mathbf{x}) = \prod_{i=1}^{d} p(x_i | x_{i+1}, ..., x_d)$$

$$p(\mathbf{x}) = \prod_{i=1}^{d} p(x_i | \mathbf{PA}_i)$$

$$\frac{d\mathbf{x}}{dt} = f(\mathbf{x}), \ \mathbf{x} \in \mathbb{R}^d$$

$$\mathbf{x}(t_0) = \mathbf{x}_0$$

Statistical

- IID
- Easiest to learn
- Statistical dependencies

Causal

- IID/non-IID (interventional)
- Harder to learn
- Statistical dependencies
- Causal structure
- Effect of interventions

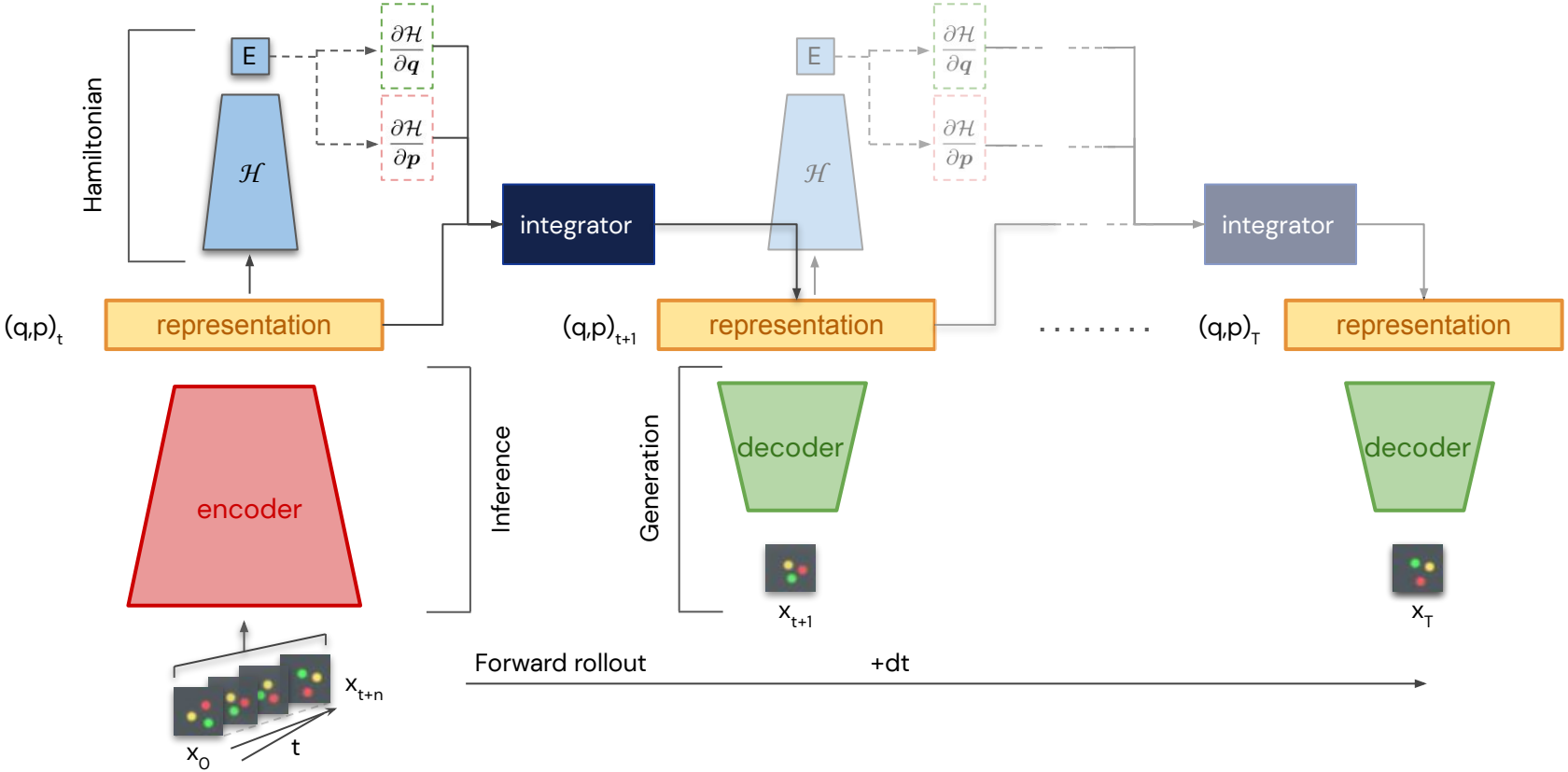
Physical

- Non-IID (arrow of time)
- Hardest to learn
- Statistical dependencies
- Causal structure
- Effect of interventions
- Future state of system
- Full description

Hamiltonian Generative Network (HGN)

Want to learn more?

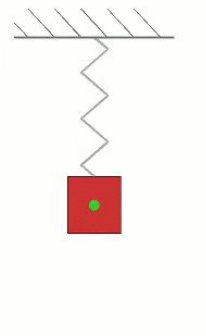
Hamiltonian Generative Networks, Toth, Rezende et al, ICLR 2020

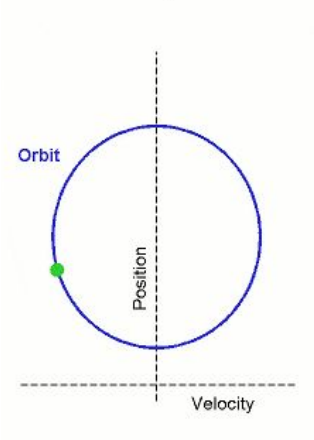


What is a Hamiltonian?

Want to learn more?

Hamiltonian Generative Networks, Toth, Rezende et al, ICLR 2020





q: position

p: momentum

phase space

Hamiltonian: $$H(q,p) = \frac{1}{2}kq^2 + \frac{p^2}{2m}$$

Time evolution: $$\frac{dq}{dt} = \frac{\partial H}{\partial p} \qquad \frac{dp}{dt} = -\frac{\partial H}{\partial q}$$

E.g. with Euler integrator:

$$q_{t+1} = q_t + dt\ \frac{\partial H}{\partial p} \qquad p_{t+1} = p_t - dt\ \frac{\partial H}{\partial q}$$

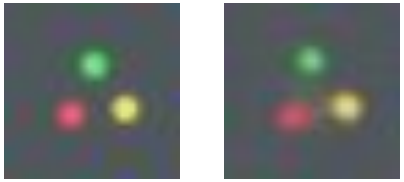
Power of Physical modeling

Want to learn more?

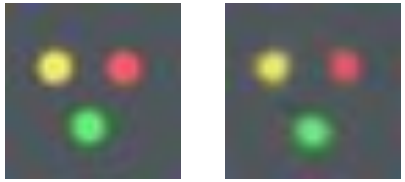
Hamiltonian Generative Networks, Toth, Rezende et al, ICLR 2020

Original Reconstructed Reversed

Original 2x slower



Original 2x faster



Samples

Sample 1 Sample 2 Sample 3 Sample 4 Sample 5 Sample 6

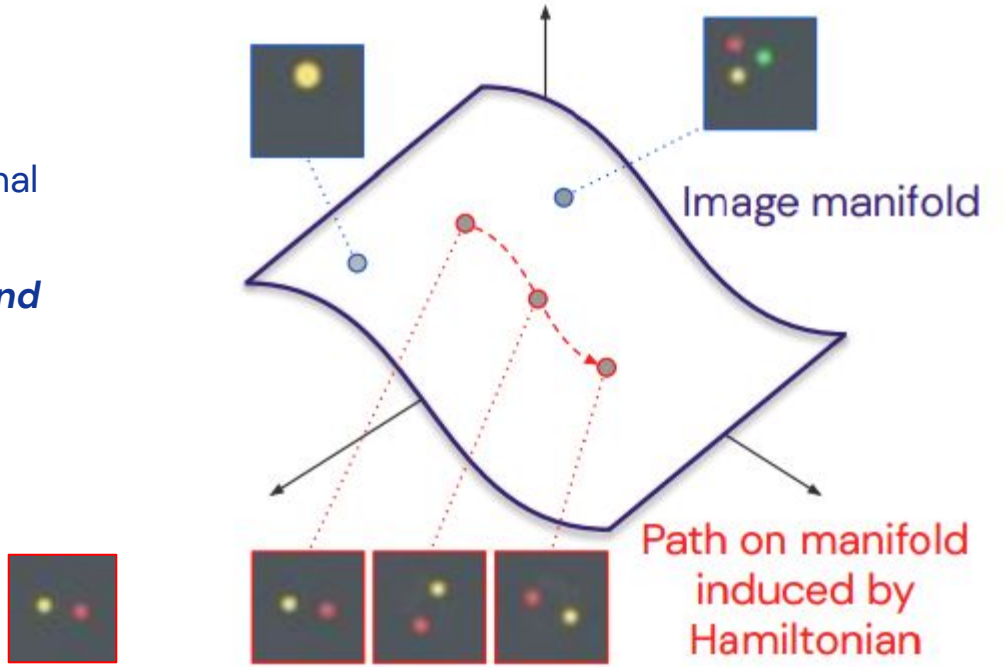
Hamiltonian Manifold Hypothesis

Want to learn more?

Hamiltonian Generative Networks, Toth, Rezende et al, ICLR 2020

Natural images lie on a low dimensional manifold in pixel space and

natural image sequences correspond to motion governed by abstract Hamiltonian dynamics.



Levels of modelling

Want to learn more?

Causality for Machine Learning, Schölkopf, arxiv 2019

$$p(\mathbf{x}) = \prod_{i=1}^{d} p(x_i | x_{i+1}, ..., x_d)$$

$$p(\mathbf{x}) = \prod_{i=1}^{d} p(x_i | \mathbf{PA}_i)$$

$$\frac{d\mathbf{x}}{dt} = f(\mathbf{x}), \ \mathbf{x} \in \mathbb{R}^d$$

$$\mathbf{x}(t_0) = \mathbf{x}_0$$

Statistical

- IID
- Easiest to learn
- Statistical dependencies

Causal

- IID/non-IID (interventional)
- Harder to learn
- Statistical dependencies
- Causal structure
- Effect of interventions

Physical

- Non-IID (arrow of time)
- Hardest to learn
- Statistical dependencies
- Causal structure
- Effect of interventions
- Future state of system
- Full description

Statistical perspective of disentangling

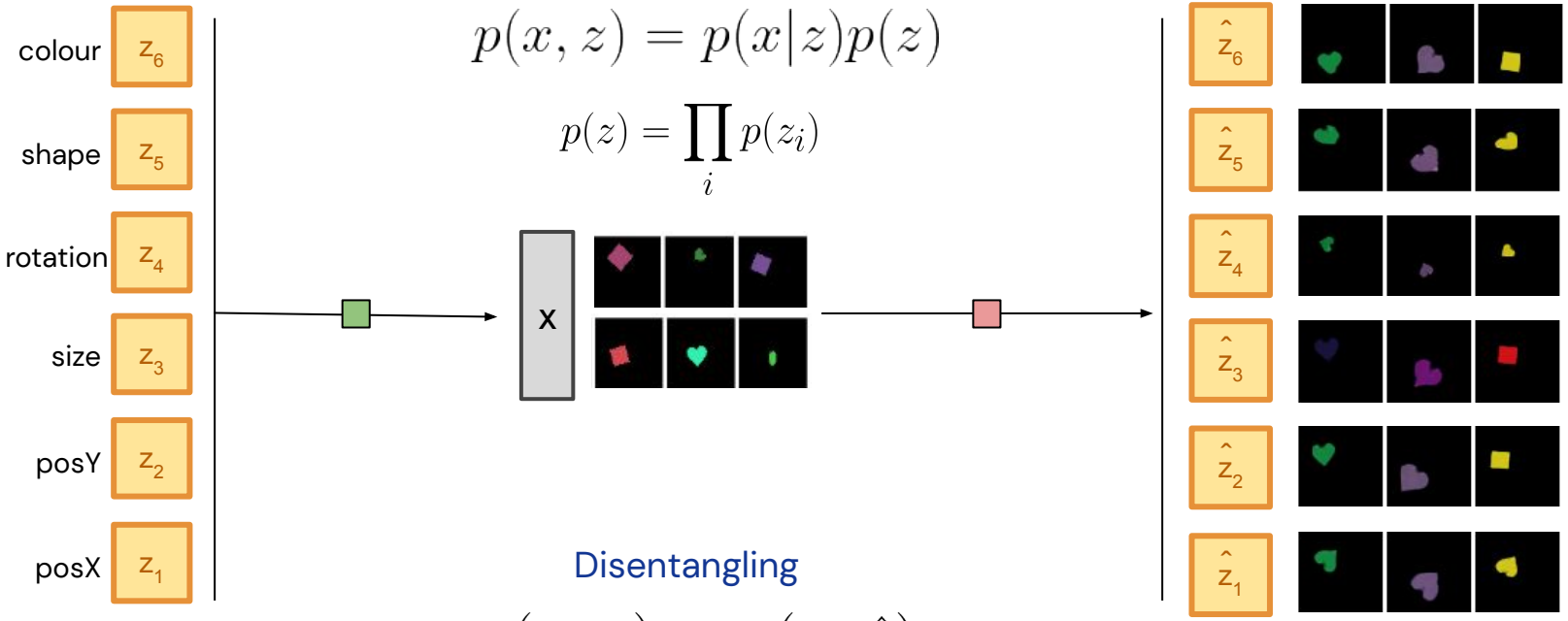
Want to learn more?

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

Generative model

$$p(x, z) = p(x|z)p(z)$$

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



Disentangling

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

GIFs adapted from Chris Burgess

Statistical perspective of disentangling

Want to learn more?

Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations, Locatello et al, ICML 2019

Theorem 1. *For $d > 1$, let $\mathbf{z} \sim P$ denote any distribution which admits a density $p(\mathbf{z}) = \prod_{i=1}^{d} p(\mathrm{z}_i)$. Then, there exists an infinite family of bijective functions $f : \mathrm{supp}(\mathbf{z}) \to \mathrm{supp}(\mathbf{z})$ such that $\frac{\partial f_i(\mathbf{u})}{\partial u_j} \neq 0$ almost everywhere for all i and j (i.e., $\mathbf{z}$ and $f(\mathbf{z})$ are completely entangled) and $P(\mathbf{z} \leq \mathbf{u}) = P(f(\mathbf{z}) \leq \mathbf{u})$ for all $\mathbf{u} \in \mathrm{supp}(\mathbf{z})$ (i.e., they have the same marginal distribution).*

- Disentangled representations are **non-identifiable** in naive unsupervised setting
- Inductive **biases in models and/or data** make unsupervised disentangling work in practice (e.g. β-VAE, FactorVAE, TC-VAE etc)

Statistical perspective of disentangling

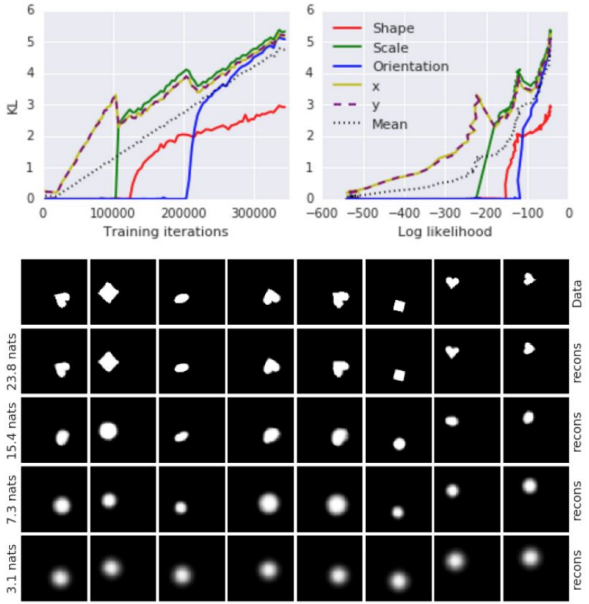
Want to learn more?

Understanding Disentangling in β-VAE. Burgess et al, arxiv 2018

Variational Autoencoders Pursue PCA Directions (by Accident). Rolinek, Zietlow et al, CVPR 2019

Theorem 1. *For $d > 1$, let $\mathbf{z} \sim P$ denote any distribution which admits a density $p(\mathbf{z}) = \prod_{i=1}^{d} p(\mathrm{z}_i)$. Then, there exists an infinite family of bijective functions $f : \mathrm{supp}(\mathbf{z}) \to \mathrm{supp}(\mathbf{z})$ such that $\frac{\partial f_i(\mathbf{u})}{\partial u_j} \neq 0$ almost everywhere for all i and j (i.e., $\mathbf{z}$ and $f(\mathbf{z})$ are completely entangled) and $P(\mathbf{z} \leq \mathbf{u}) = P(f(\mathbf{z}) \leq \mathbf{u})$ for all $\mathbf{u} \in \mathrm{supp}(\mathbf{z})$ (i.e., they have the same marginal distribution).*

- Disentangled representations are **non-identifiable** in naive unsupervised setting
- Inductive **biases in models and/or data** make unsupervised disentangling work in practice (e.g. β-VAE, FactorVAE, TC-VAE etc)



Optimizing the stochastic part of the reconstruction loss promotes local orthogonality of the decoder.

Levels of modelling

Want to learn more?

Causality for Machine Learning, Schölkopf, arxiv 2019

$$p(\mathbf{x}) = \prod_{i=1}^{d} p(x_i | x_{i+1}, ..., x_d)$$

$$p(\mathbf{x}) = \prod_{i=1}^{d} p(x_i | \mathbf{PA}_i)$$

$$\frac{d\mathbf{x}}{dt} = f(\mathbf{x}), \ \mathbf{x} \in \mathbb{R}^d$$

$$\mathbf{x}(t_0) = \mathbf{x}_0$$

Statistical

- IID
- Easiest to learn
- Statistical dependencies

Causal

- IID/non-IID (interventional)
- Harder to learn
- Statistical dependencies
- Causal structure
- Effect of interventions

Physical

- Non-IID (arrow of time)
- Hardest to learn
- Statistical dependencies
- Causal structure
- Effect of interventions
- Future state of system
- Full description

Structural causal models (SCMs)

Want to learn more?

Causality for Machine Learning, Schölkopf, arxiv 2019

Given set of observables $x_1, \ldots, x_n$ (random variables, vertices of a DAG), each can be expressed as:

$$x_i := f_i(\mathbf{PA}_i, U_i)$$

$\mathbf{PA}_i$ - parents

f_i - deterministic function

U_i - stochastic unexplained variable, where $U_1, \ldots, U_n$ are jointly independent

Given SCM can express conditional probabilities $p(x_i|\mathbf{PA}_i)$

Independent causal mechanisms

Want to learn more?

Causality for Machine Learning, Schölkopf, arxiv 2019

Independent Causal Mechanisms (ICM) Principle. *The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other.*

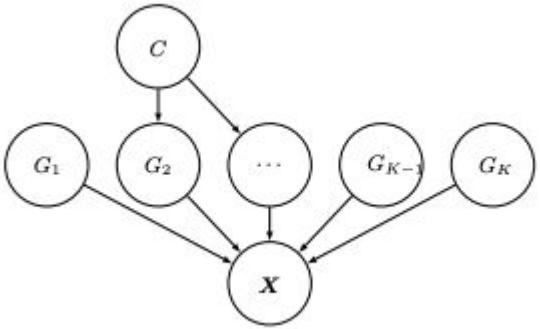
In the probabilistic case, this means that the conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other mechanisms.

- Changing (or intervening upon) one mechanism $p(x_i|\mathbf{PA}_i)$ does not change other mechanisms $p(x_j|\mathbf{PA}_j),\ j \neq i$
- Knowing about other mechanisms $p(x_j|\mathbf{PA}_j),\ j \neq i$ does not inform about mechanism $p(x_i|\mathbf{PA}_i)$

Causal view of disentangling

Want to learn more?

Robustly Disentangled Causal Mechanisms: Validating Deep Representations for Interventional Robustness, Suter et al, ICML 2019



Definition 1 (Disentangled Causal Process). *Consider a causal model for $\boldsymbol{X}$ with generative factors $\boldsymbol{G}$, described by the mechanisms $p(\boldsymbol{x}|\boldsymbol{g})$, where $\boldsymbol{G}$ could generally be influenced by L confounders $\boldsymbol{C} = (C_1, \dots, C_L)$. This causal model for $\boldsymbol{X}$ is called disentangled if and only if it can be described by a structural causal model (SCM) (Pearl, 2009) of the form*

$$\boldsymbol{C} \leftarrow \boldsymbol{N}_c$$
$$G_i \leftarrow f_i(\boldsymbol{PA}_i^C, N_i), \quad \boldsymbol{PA}_i^C \subset \{C_1, \dots, C_L\},\ i = 1, \dots, K$$
$$\boldsymbol{X} \leftarrow g(\boldsymbol{G}, N_x)$$

with functions f_i, g and jointly independent noise variables $\boldsymbol{N}_c, N_1, \dots, N_K, N_x$. Note that $\forall i \neq j\ G_i \not\to G_j$.

Proposition 1 (Properties of a Disentangled Causal Process). *A disentangled causal process as introduced in Definition 1 fulfills the following properties:*

(a) $p(\boldsymbol{x}|\boldsymbol{g})$ describes a causal mechanism invariant to changes in the distributions $p(g_i)$.

(b) In general, the latent causes can be dependent

$$G_i \not\perp\!\!\!\perp G_j, i \neq j.$$

Only if we condition on the confounders in the data generation they are independent

$$G_i \perp\!\!\!\perp G_j | \boldsymbol{C} \quad \forall i \neq j.$$

(c) Knowing what observation of $\boldsymbol{X}$ we obtained renders the different latent causes dependent, i.e.,

$$G_i \not\perp\!\!\!\perp G_j | \boldsymbol{X}.$$

(d) The latent factors $\boldsymbol{G}$ already contain all information about confounders $\boldsymbol{C}$ that is relevant for $\boldsymbol{X}$, i.e.,

$$I(\boldsymbol{X}; \boldsymbol{G}) = I(\boldsymbol{X}; (\boldsymbol{G}, C)) \geq I(\boldsymbol{X}; C)$$

where I denotes the mutual information.

(e) There is no total causal effect from G_j to G_i for $j \neq i$; i.e., intervening on G_j does not change G_i, i.e,

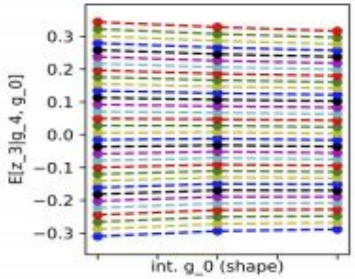
$$\forall g_j^{\triangle} \quad p(g_i | do(G_j \leftarrow g_j^{\triangle})) = p(g_i) \quad \left(\neq p(g_i | g_j^{\triangle})\right)$$

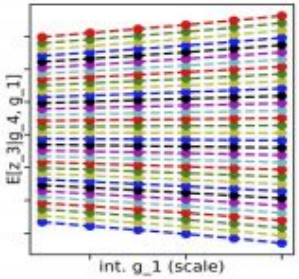
Measuring disentangling through causality

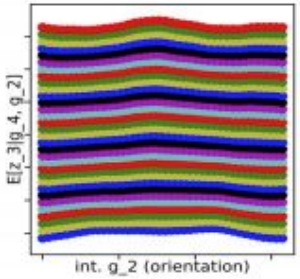
Want to learn more?

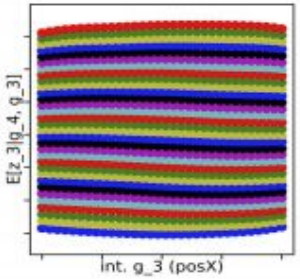
Robustly Disentangled Causal Mechanisms: Validating Deep Representations for Interventional Robustness, Suter et al, ICML 2019

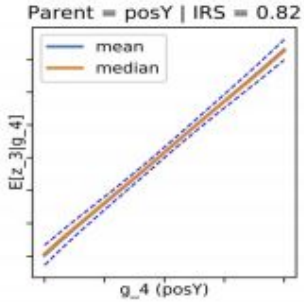
Well disentangled according to causal metric (IRS)



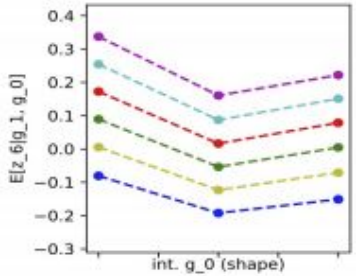


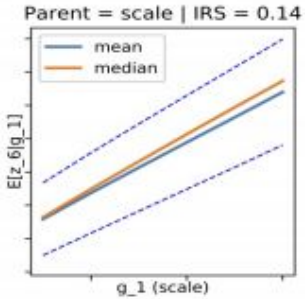


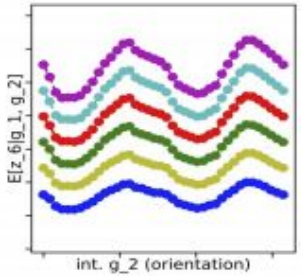


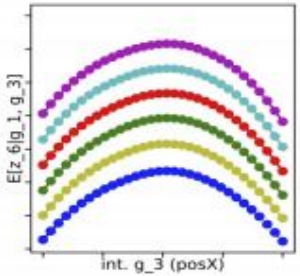


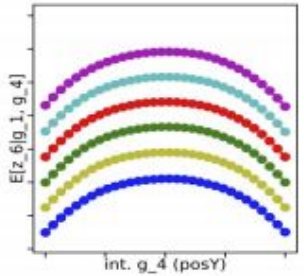
Badly disentangled according to causal metric (IRS)

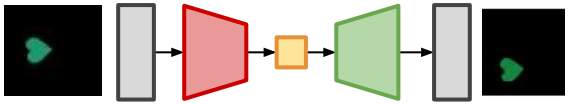












Levels of modelling

Want to learn more?

Causality for Machine Learning, Schölkopf, arxiv 2019

$$p(\mathbf{x}) = \prod_{i=1}^{d} p(x_i | x_{i+1}, ..., x_d)$$

$$p(\mathbf{x}) = \prod_{i=1}^{d} p(x_i | \mathbf{PA}_i)$$

$$\frac{d\mathbf{x}}{dt} = f(\mathbf{x}), \ \mathbf{x} \in \mathbb{R}^d$$

$$\mathbf{x}(t_0) = \mathbf{x}_0$$

Statistical

- IID
- Easiest to learn
- Statistical dependencies

Causal

- Non-IID
- Harder to learn
- Statistical dependencies
- Causal structure
- Effect of interventions

Physical

- Non-IID (arrow of time)
- Hardest to learn
- Statistical dependencies
- Causal structure
- Effect of interventions
- Future state of system
- Full description

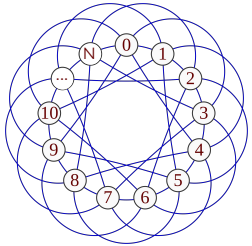
Physical definition of disentangling

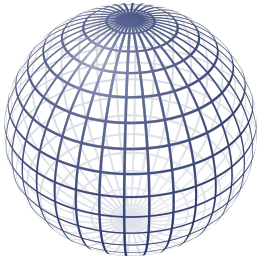
Want to learn more?

Towards a Definition of Disentangled Representations, Higgins, Amos et al, ICML Workshop on Theoretical Physics for Deep Learning 2019

A vector representation is called a ***disentangled representation*** *with respect to a particular decomposition of a symmetry group into subgroups, if it decomposes into independent subspaces, where each subspace is affected by the action of a single subgroup, and the actions of all other subgroups leave the subspace unaffected.*

Primer on group theory: examples





$$R_z(\varphi) = \begin{bmatrix} \cos\varphi & -\sin\varphi & 0 \\ \sin\varphi & \cos\varphi & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Cyclic group – C_N

Group generated by a single element g. Every element of the group may be obtained by repeatedly applying the group operation to g or its inverse.

Rotational symmetries of a polygon forms a finite cyclic group.

Describes integers modulo N.

3D rotation group – SO(3)

Group of all rotations about the origin of 3D Euclidean space R^3 under the operation of composition.

Group of all orthogonal 3x3 matrices with determinant 1.

Describes rotational symmetries of 3D objects, orientations.

Primer on group theory: group and group action

A **group** is a set G with an operator $\circ$: $G \times G \rightarrow G$:

$$G = \{ \{e, g_1, g_2\}, \circ \}$$

A group has four properties:

1) Associativity $\forall\ x,y,z \in G : x \circ (y \circ z) = (x \circ y) \circ z$
2) Identity $\exists\ e \in G\ \forall\ x \in G : e \circ x = x \circ e = x$
3) Inverse $\forall\ x \in G\ \exists\ x^{-1} \in G : x \circ x^{-1} = x^{-1} \circ x = e$

Primer on group theory: group action

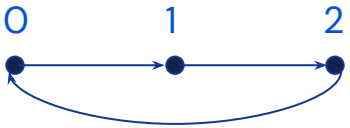
Group **action** on (set) W:

$\cdot : G \times W \to W$

should satisfy:

1) Identity $ew = w \quad \forall\ w \in W$
2) Associativity $(gh)w = g(hw) \quad \forall\ g, h \in G,\ w \in W$

$W = \{ [0], [1], [2] \}$

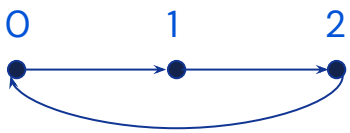


	e	$\mathbf{g_1}$	$\mathbf{g_2}$
e	e	g_1	g_2
$\mathbf{g_1}$	g_1	g_2	e
$\mathbf{g_2}$	g_2	e	g_1

Primer on group theory: subgroup

$G = \{ \{e, g_1, g_2\}, \circ \}$

$W = \{ [0], [1], [2] \}$



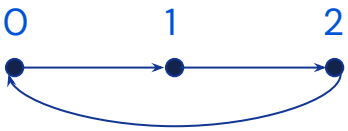
	e	$\mathbf{g_1}$	$\mathbf{g_2}$
e	e	g_1	g_2
$\mathbf{g_1}$	g_1	g_2	e
$\mathbf{g_2}$	g_2	e	g_1

$G = C_3$

Primer on group theory: subgroup

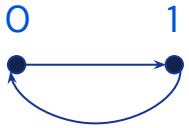
$G = \{ \{e, g_1, g_2\}, \circ \}$

$W = \{ [0], [1], [2] \}$



H	**e**	**g_1**	g_2
e	e	g_1	g_2
g_1	g_1	e	e
g_2	g_2	e	g_1

$G = C_3$ $H = C_2$



H is a **subgroup** of G, if it is a subset of G, that is closed under $\circ$ operator and inverses:

$x,y \in H \Rightarrow x \circ y \in H \wedge x^{-1} \in H$

Primer on group theory: direct product

Given two groups G and H, can construct a new group K = G x H (**direct product**):

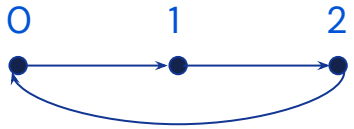
1) Underlying set is a cartesian product G x H, i.e. ordered pairs (g, h) where $g \in G$ and $h \in H$

2) Group operator $\bullet$ defined component wise, i.e. $(g_1, h_1) \bullet (g_2, h_2) = (g_1 \circ g_2, h_1 \odot h_2)$

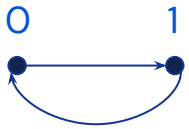
$K = G \times H$

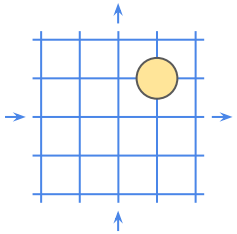
G \ H	$e_g e_h$	$g_1 e_h$	$g_2 e_h$	$e_g h_1$	$g_1 h_1$	$g_2 h_1$
$e_g e_h$	$e_g e_h$	$g_1 e_h$	$g_2 e_h$	$e_g h_1$	...	...
$g_1 e_h$	$g_1 e_h$	$g_2 e_h$	$e_g e_h$	...	...	...
$g_2 e_h$	$g_2 e_h$	$e_g e_h$	$g_1 e_h$	...	...	...
$e_g h_1$	$e_g h_1$	...	...	$e_g e_h$	...	...
$g_1 h_1$	...	...	...	...	...	...
$g_2 h_1$	...	...	...	...	...	...

$G = \{ \{e_g, g_1, g_2\}, \circ \}$



$H = \{ \{e_h, h_1\}, \odot \}$





Position y

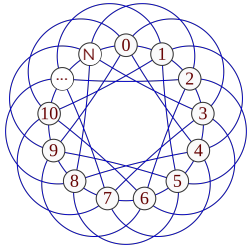
Position x

Colour

$G_x = C_N$

$G_y = C_N$

$G_c = C_N$



Want to learn more?

Towards a Definition of Disentangled Representations, Higgins, Amos et al, ICML Workshop on Theoretical Physics for Deep Learning 2019

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

W

(**x=5**, y=6, c=yellow)

b

O

h

Z

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

f

$\cdot : G \times W \to W$

Want to find an **equivariant map** f, s.t.:

$g \cdot f(w) = f(g \cdot w) \quad \forall \; g \in G, w \in W$

$* : G \times Z \to Z$

W

(**x=3**, y=6, c=yellow)

b

O

h

Z

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

f

Summary

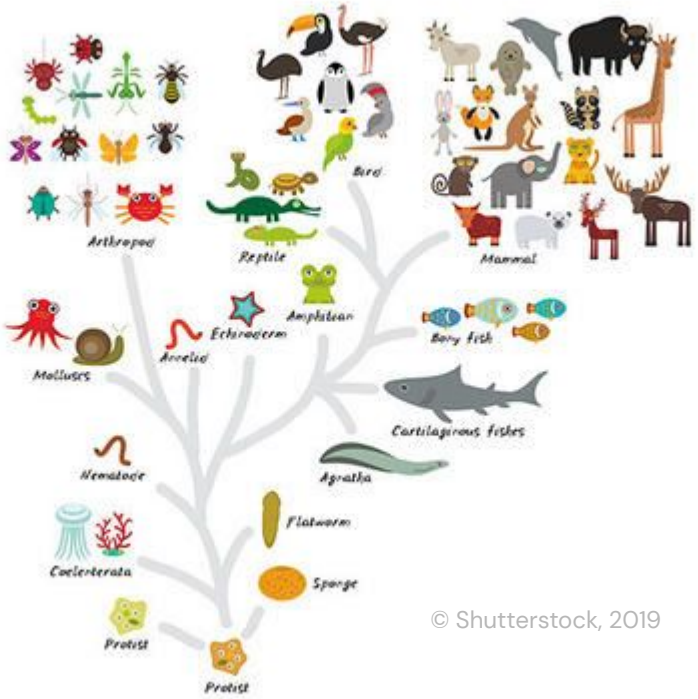
The diversity of models for unsupervised representation learning appears large – the “model zoo”

The large space of models can be understood and navigated using a relatively simple theoretical taxonomy:

1. How is the **data density** modelled?
2. What properties of the **manifold** are modelled?
3. What **level of modeling detail** is expected?

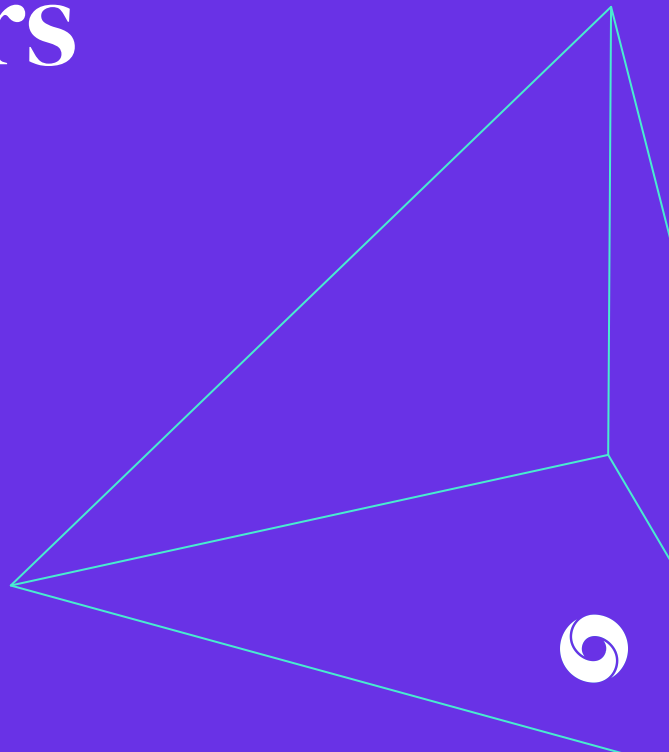
Thinking about these questions may help understand:

1. the **tradeoffs** of different implementational choices
2. **troubleshoot** failures
3. possible directions of model **improvement**



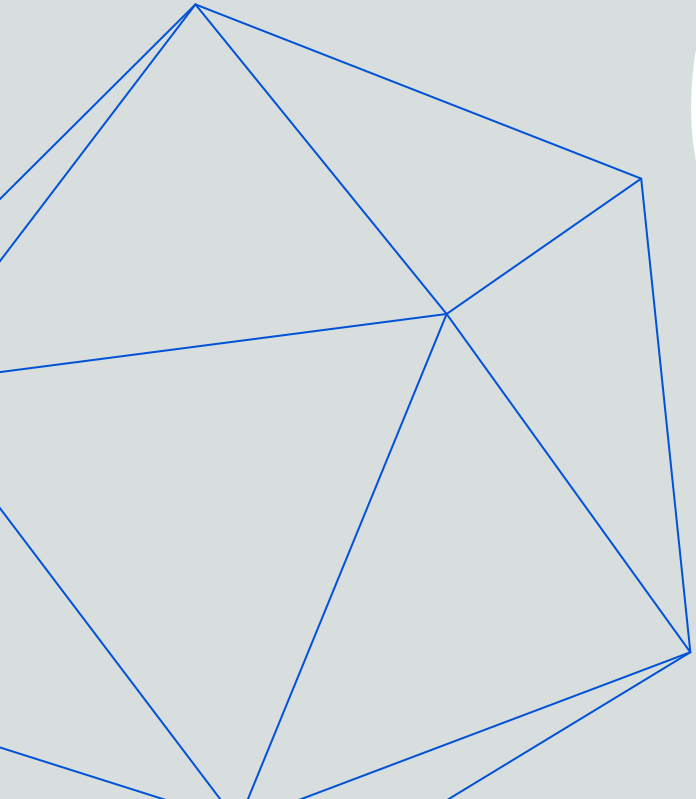
DeepMind

4 Frontiers



DeepMind

Multi-disciplinary perspective



Sources of inspiration

linguistics

physics

machine learning

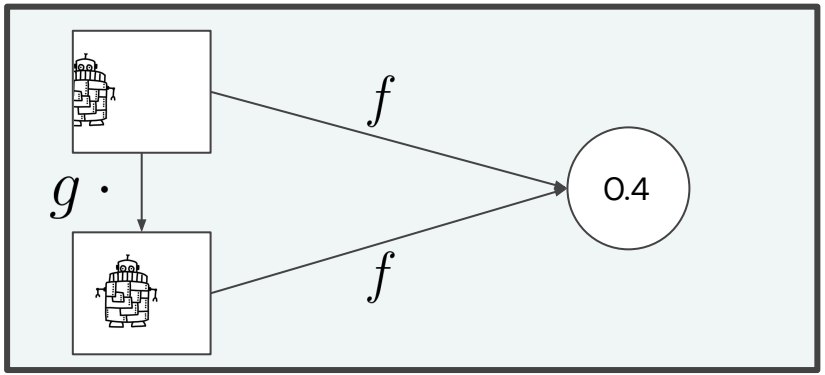
neuroscience

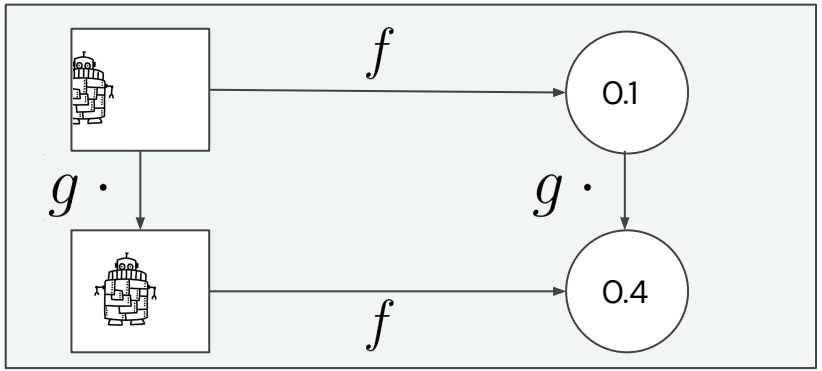
mathematics

Invariance vs equivariance

Want to learn more?

Backpropagation Applied to Handwritten Zip Code Recognition, LeCun et al, Neural Computation 1989





Invariance

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

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

Equivariance

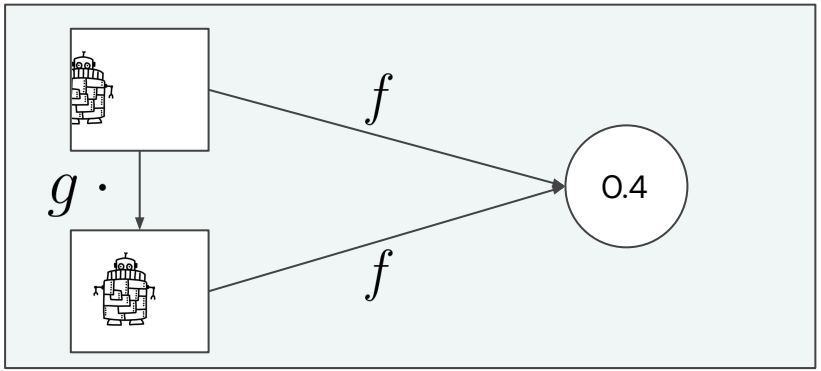
- representation reflects the transformation applied to the input

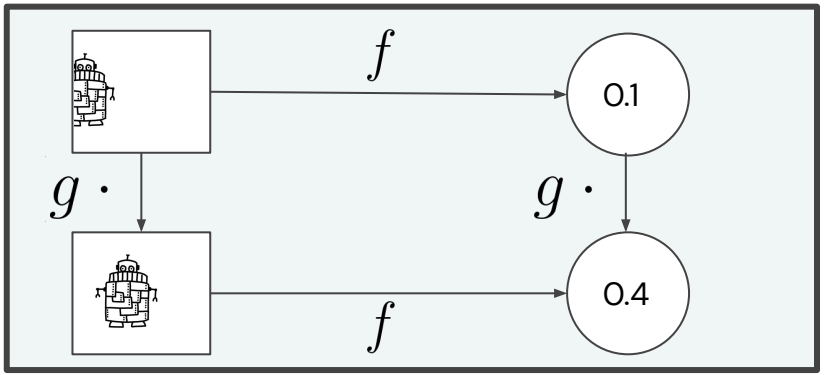
$$f(g \cdot x) = g \cdot f(x)$$

Invariance vs equivariance

Want to learn more?

A General Theory of Equivariant CNNs on Homogeneous Spaces, Cohen et al, NeurIPS 2019





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

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

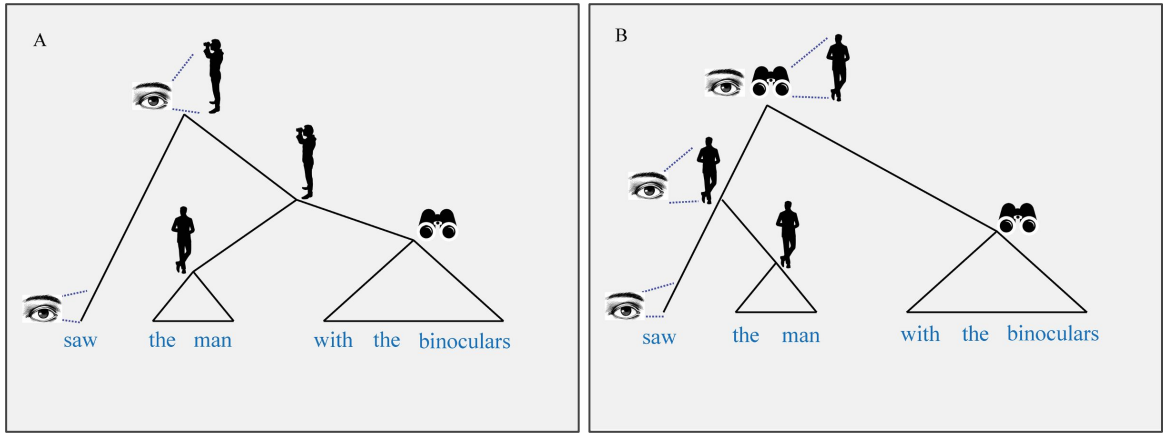
Compositionality

Want to learn more?

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

"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*.



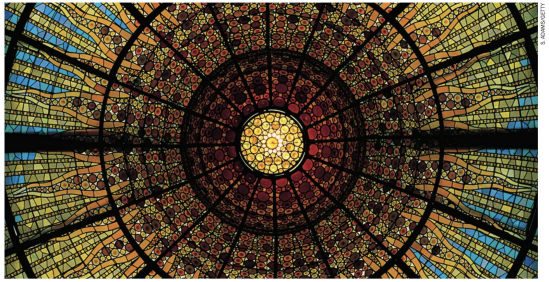
Bolhuis et al, 2018

Symmetry transformations

Want to learn more?

Why symmetry matters, Livio, Nature 2012

COMMENT



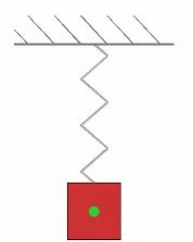
Symmetries feature in the stained-glass ceiling of the Palace of Catalan Music in Barcelona, Spain.

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

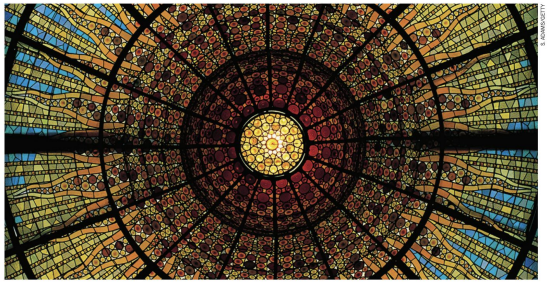


Symmetry transformations

Want to learn more?

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

COMMENT



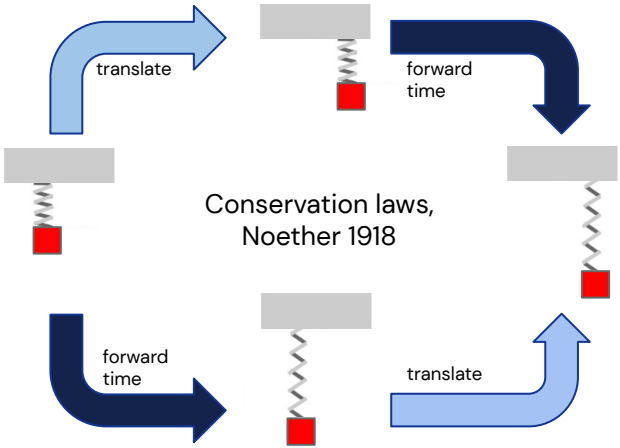
Symmetries feature in the stained-glass ceiling of the Palace of Catalan Music in Barcelona, Spain.

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

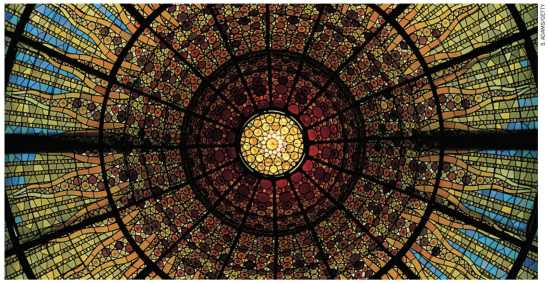


Studying symmetries of a system helps:

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

Symmetry transformations

COMMENT



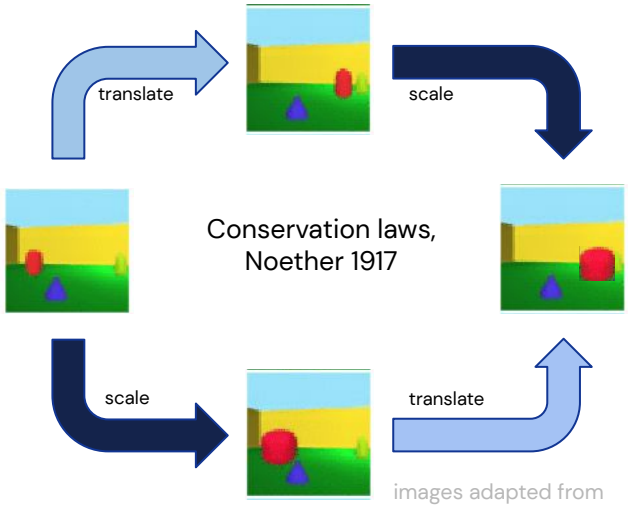
Symmetries feature in the stained-glass ceiling of the Palace of Catalan Music in Barcelona, Spain.

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



images adapted from Chris Burgess

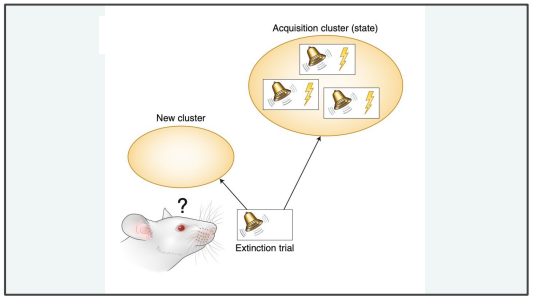
Studying symmetries of a system helps:

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

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

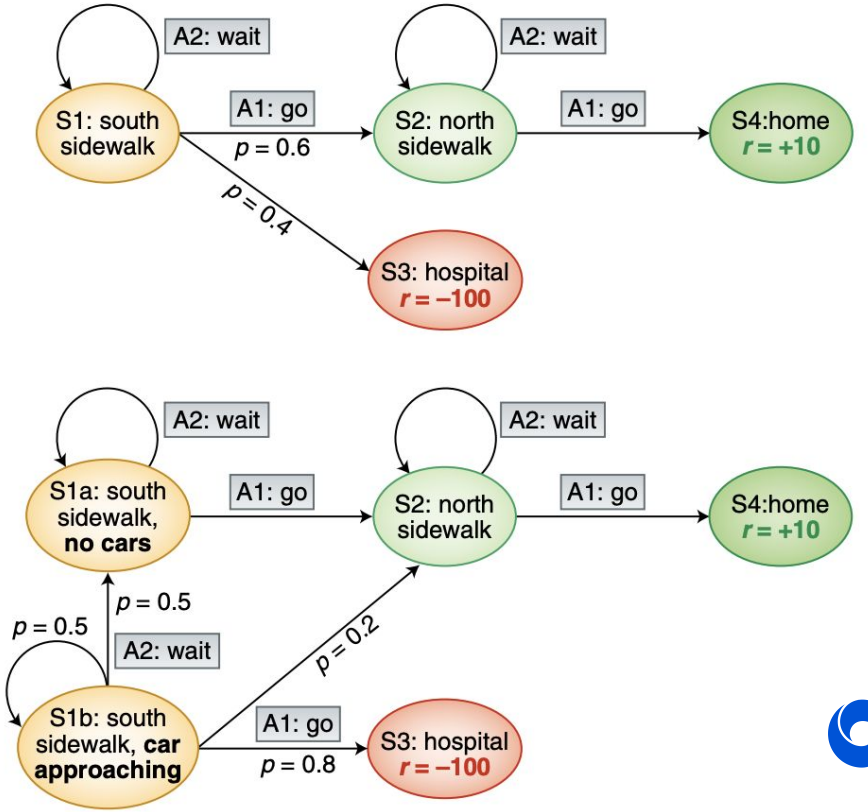
Not all information may be present in perceptual input. Representations should include information about latent aspects of the state too.


How does one cross a street?
Want to learn more?
Learning Task-State Representations, Niv, Nature Neuroscience 2019
BUSES ONLY
4PM-7PM MON THRU FRI
7AM - 7PM MON THRU FRI EXCEPT DELIVERIES THIS BLOCK
ONE WAY
W 25 ST
PIZZA
BUS ONLY

Alternative representations for the same task

Want to learn more?

Learning Task-State Representations, Niv, Nature Neuroscience 2019

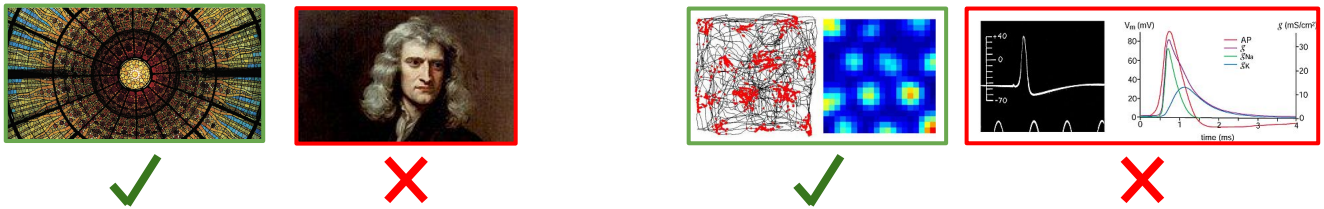


Guiding principles

Want to learn more?

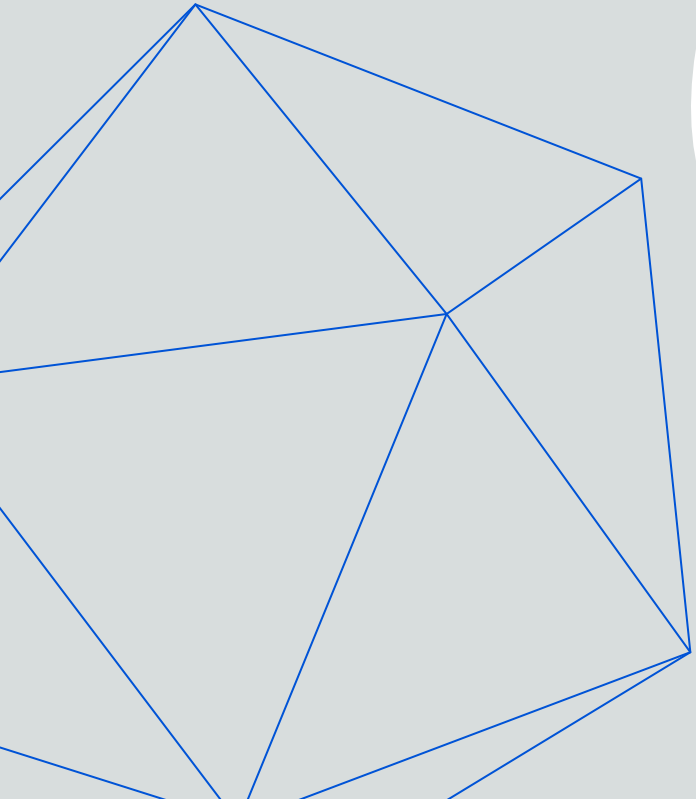
The Bitter Lesson, Sutton, incompleteideas.net 2019

Avoid details, think about the fundamentals



DeepMind

Evaluating representations

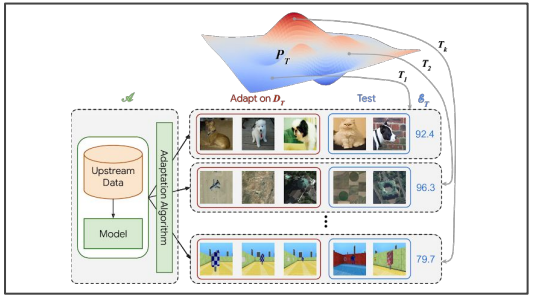


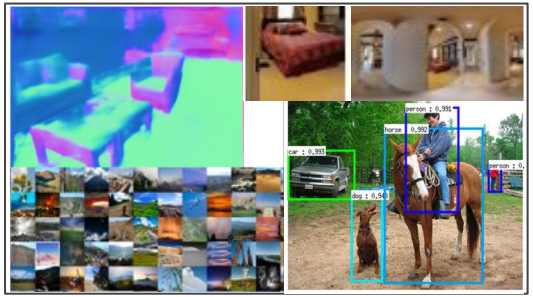
Evaluating representations

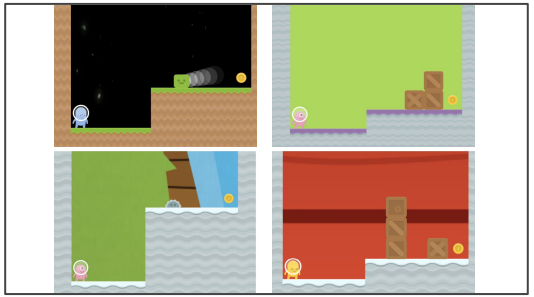
Want to learn more?

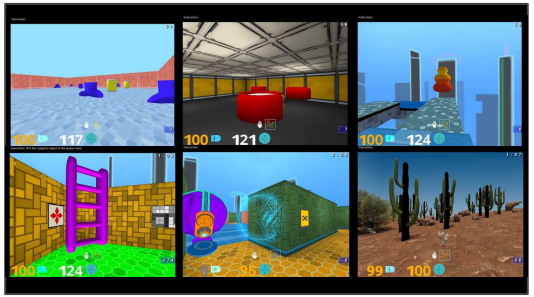
A Large-scale Study of Representation Learning with the Visual Task Adaptation Benchmark, Zhai et al, arxiv 2019

- No standardised methodology
- Based on task performance
  - Does representation help with many tasks?
  - Is task learning more data efficient?
- Based on assessing the representation properties







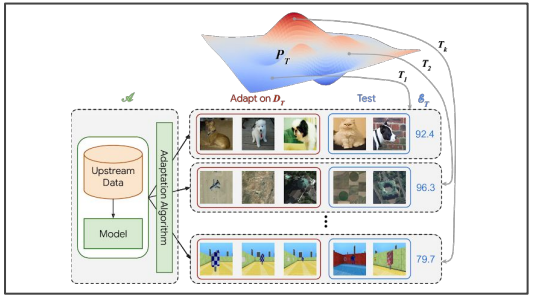


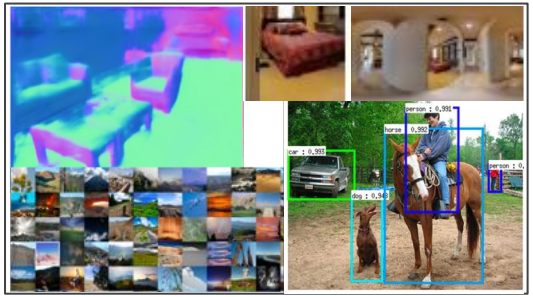
Evaluating representations

Want to learn more?

Scaling and Benchmarking Self-Supervised Visual Representation Learning, Goyal et al, ICCV 2019

- No standardised methodology
- Based on task performance
  - Does representation help with many tasks?
  - Is task learning more data efficient?





Visual Task Adaptation Benchmark (Google)

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

FAIR Self-Supervision Benchmark (Facebook)

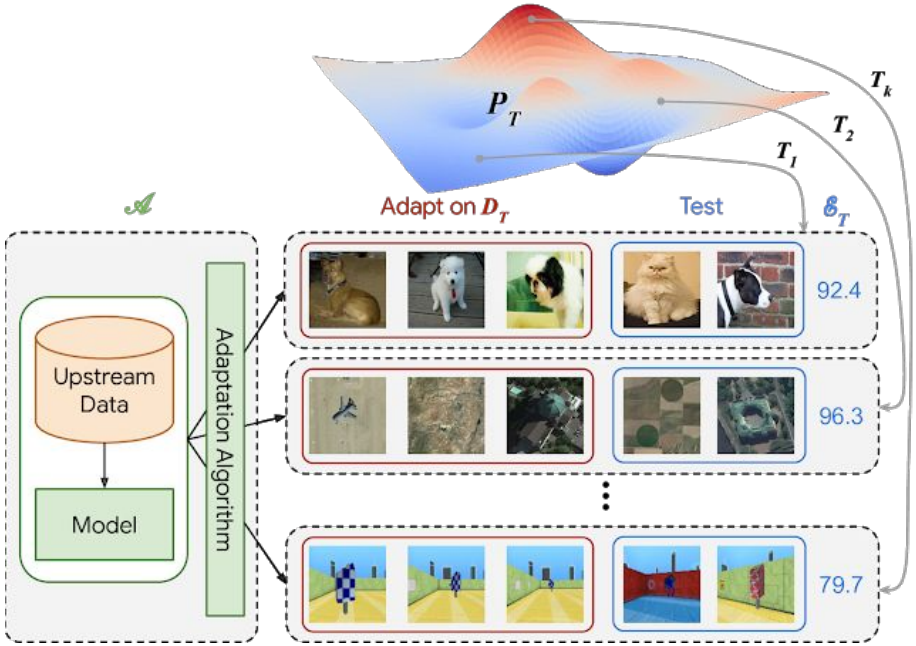
Image classification, object detection, surface normal estimation and visual navigation tasks. Allows limited supervision and fine-tuning.

Evaluating representations

Want to learn more?

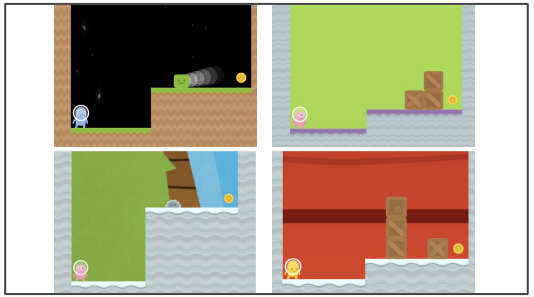
A Large-scale Study of Representation Learning with the Visual Task Adaptation Benchmark, Zhai et al, arxiv 2019

- No standardised methodology
- Based on task performance
  - Does representation help with many tasks?
  - Is task learning more data efficient?



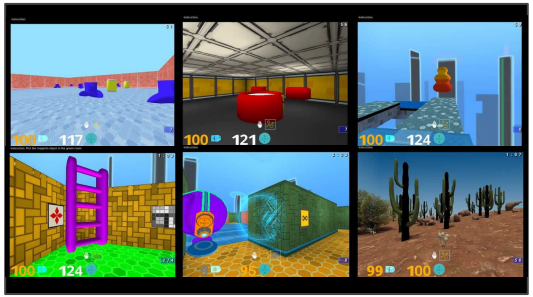
Evaluating representations

- No standardised methodology
- Based on task performance
  - Does representation help with many tasks?
  - Is task learning more data efficient?



CoinRun (OpenAI)

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



DMLab-30 (DeepMind)

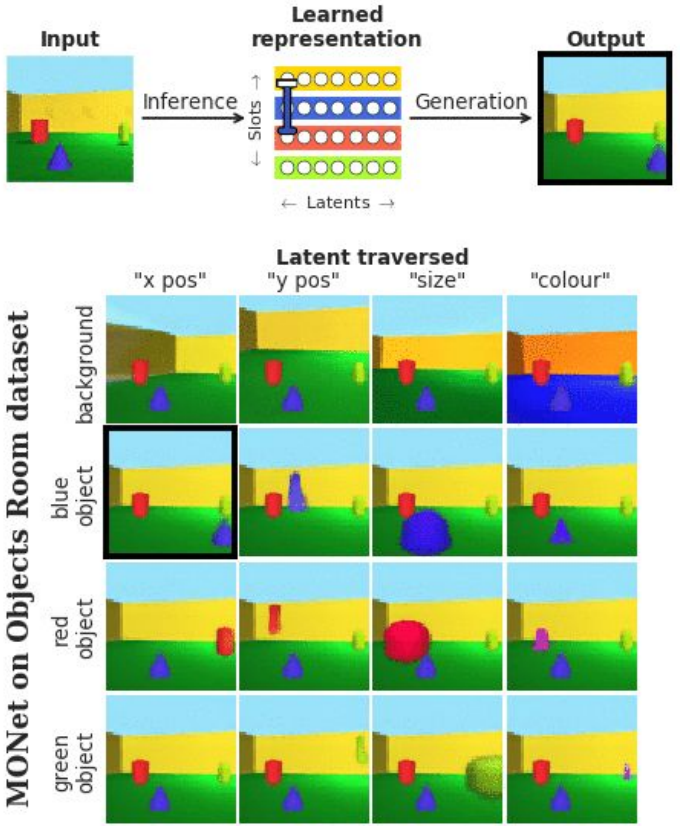
30 varied tasks in a 3D environment, testing navigation, language abilities, multi-agent interactions, long-term planning and more.

Evaluating representations

Want to learn more?

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

- No standardised methodology
- Based on task performance
  - Does representation help with many tasks?
  - Is task learning more data efficient?
- Based on assessing the representation properties
  - Latent visualisations

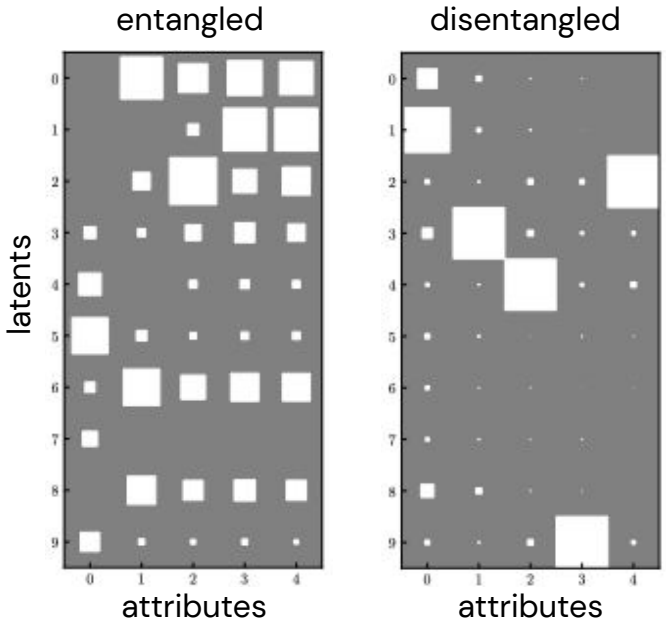


Evaluating representations

Want to learn more?

disentanglement_lib, Bachem & Locatello, github 2019
https://github.com/google-research/disentanglement_lib

- No standardised methodology
- Based on task performance
  - Does representation help with many tasks?
  - Is task learning more data efficient?
- Based on assessing the representation properties
  - Latent visualisations
  - Metrics



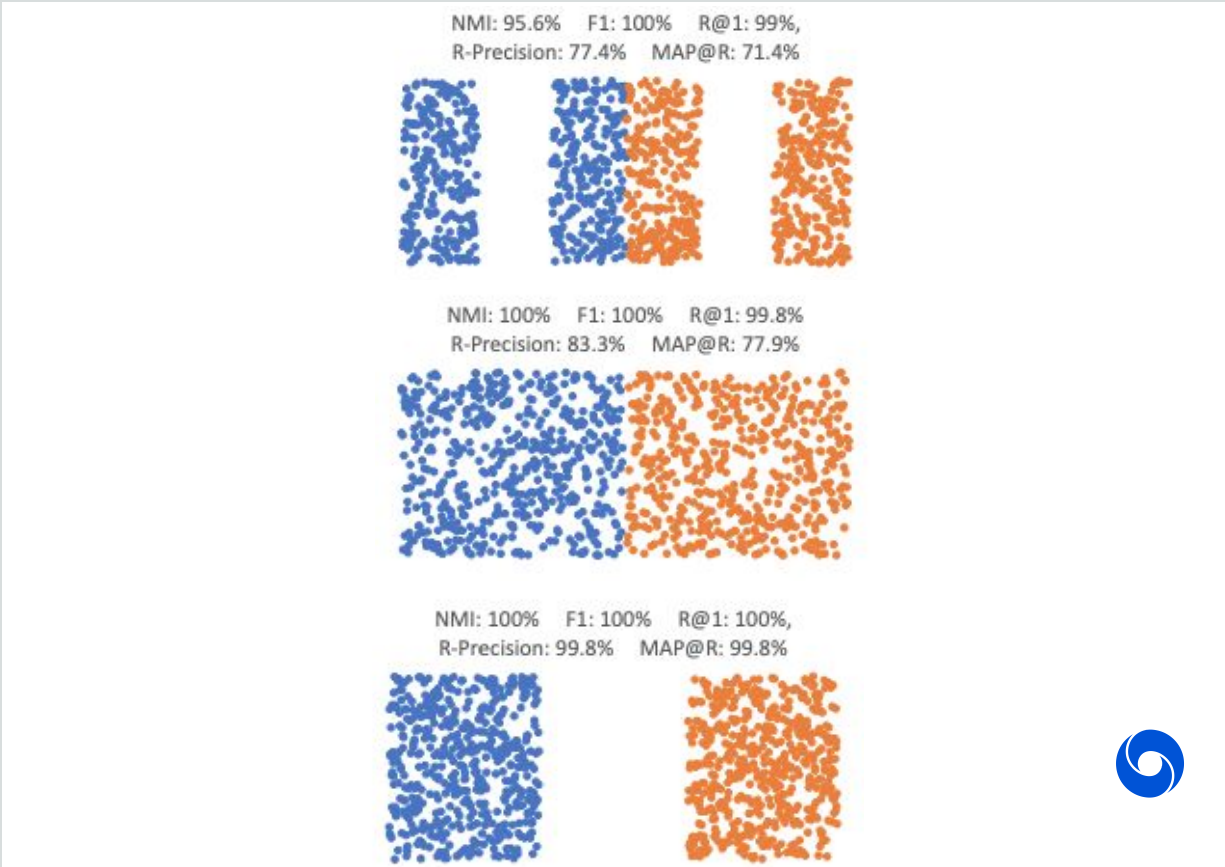
Adapted from Eastwood & Williams (2017)

Evaluating representations

Want to learn more?

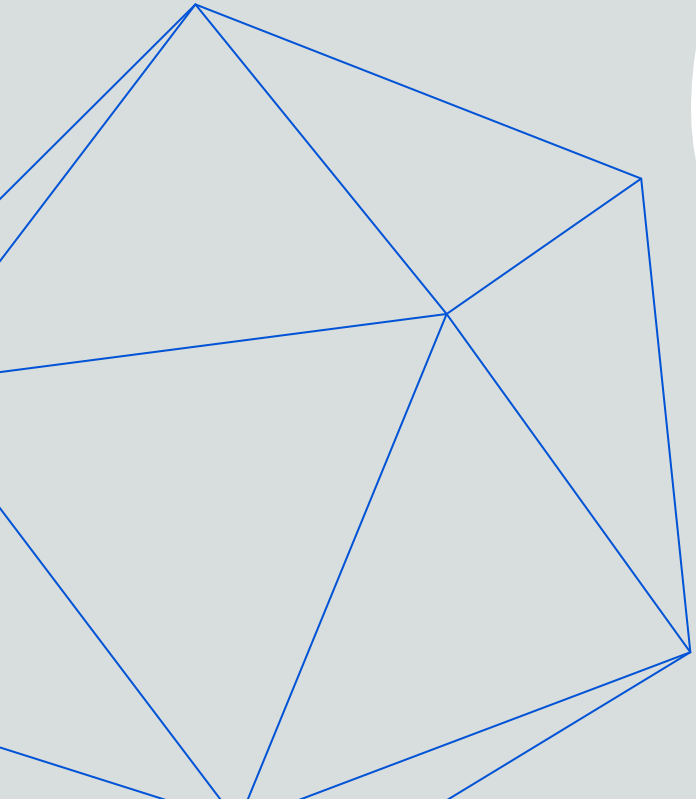
A Metric Learning Reality Check, Musgrave et al, arxiv 2020

- No standardised methodology
- Based on task performance
  - Does representation help with many tasks?
  - Is task learning more data efficient?
- Based on assessing the representation properties
  - Latent visualisations
  - Metrics



DeepMind

Utility of learning better representations



Utility of disentangled representations

Want to learn more?

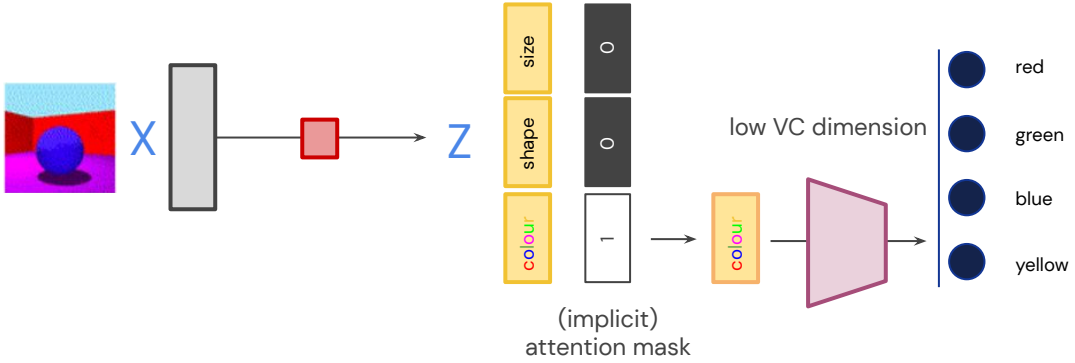
Deep Learning of Representations: Looking Forward, Bengio, SLSP 2013

- Data efficiency
- Generalisation under covariate shift
- Fairness
- Abstract reasoning

Task: what is the colour?

posX?

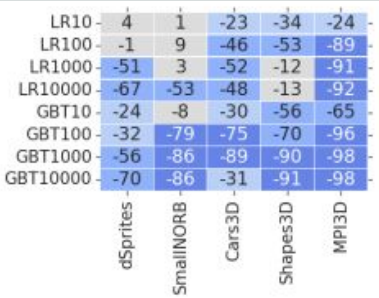
posY?



Utility of disentangled representations

Want to learn more?

Weakly-Supervised Disentanglement Without Compromises, Locatello et al, arxiv 2020

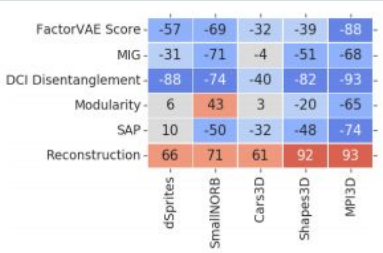


Data efficiency

Predict values of generative factors from representation in **low data regime**

Using logistic regression (LR) or gradient boosted trees (GBT)

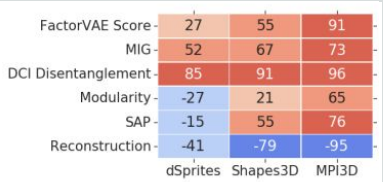
Higher disentanglement correlates with better **accuracy**



Fairness

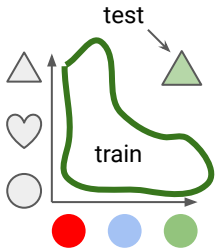
Higher disentanglement correlates with classifiers that are **fairer to unobserved sensitive variables** independent of the target variable

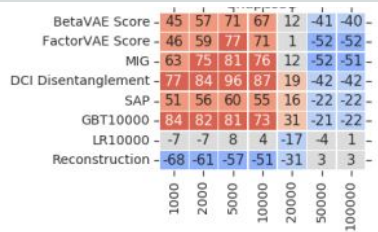
Use GBT10000 classifier



Generalisation

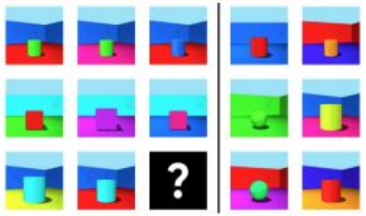
Higher disentanglement correlates with **generalisation under covariate shifts**





Abstract reasoning

Higher disentanglement correlates with **accuracy on abstract visual reasoning tasks in lower data regimes**

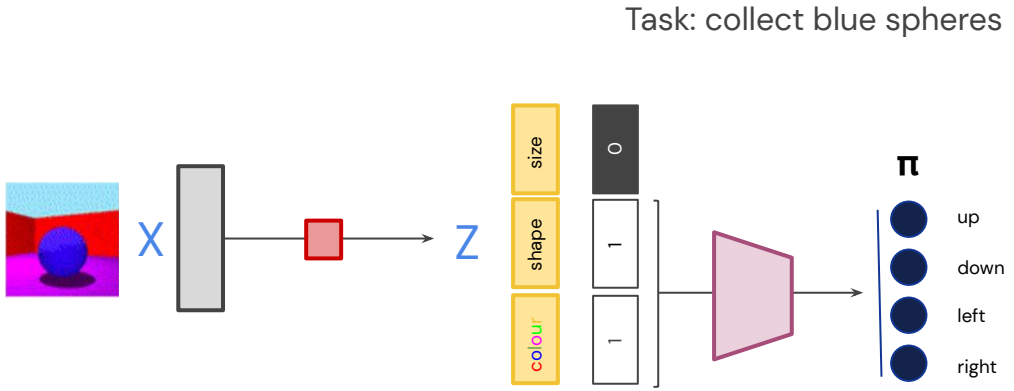


Utility of disentangled representations

Want to learn more?

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

- Data efficiency
- Generalisation under covariate shift
- Fairness
- Abstract reasoning
- Transfer

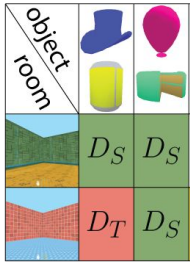


Utility of disentangled representations

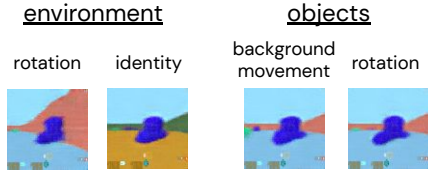
Want to learn more?

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

- Data efficiency
- Generalisation under covariate shift
- Fairness
- Abstract reasoning
- Transfer



Vision Type	DeepMind Lab DQN	DeepMind Lab A3C	DeepMind Lab EC
Baseline agent	1.86 ± 3.91	5.32 ± 3.36	-0.41 ± 4.21
UNREAL	-	4.13 ± 3.95	-
DARLA_{FT}	$\mathbf{13.36 \pm 5.8}$	1.4 ± 2.16	-
$\text{DARLA}_{\text{ENT}}$	3.45 ± 4.47	15.66 ± 5.19	5.69 ± 3.73
$\text{DARLA}_{\text{DAE}}$	7.83 ± 4.47	6.74 ± 2.81	5.59 ± 3.37
DARLA	10.25 ± 5.46	$\mathbf{19.7 \pm 5.43}$	$\mathbf{11.41 \pm 3.52}$



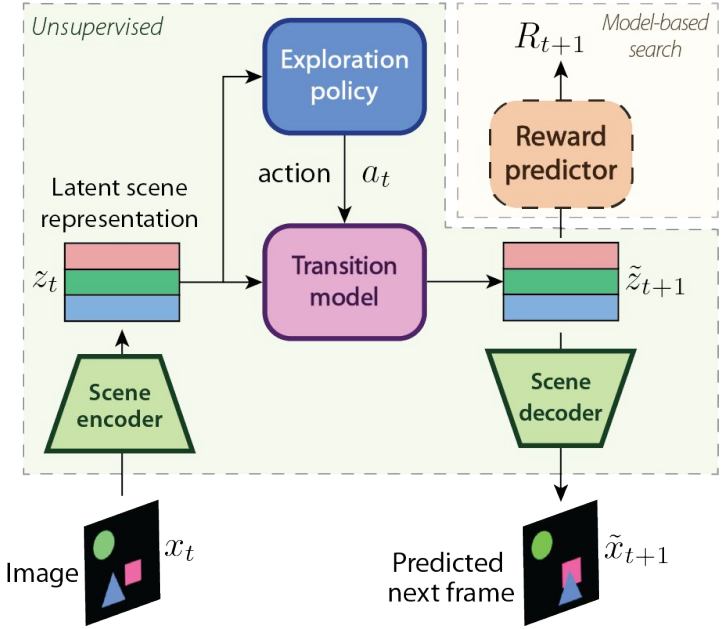
Utility of object-based representations

Want to learn more?

COBRA: Data-Efficient Model-Based RL through Unsupervised Object Discovery and Curiosity-Driven Exploration, Watters, Matthey et al, arxiv 2019

During unsupervised exploration stage learn:

- Object decomposition and feature disentangling
- Action-conditioned transition model of the environment
- Curiosity-driven exploration policy



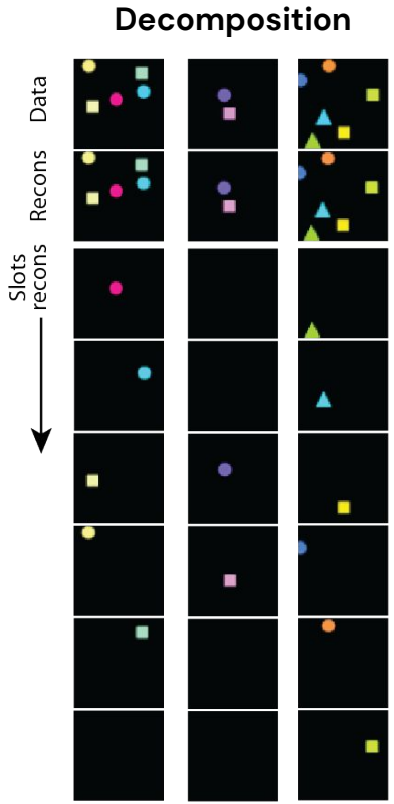
Utility of object-based representations

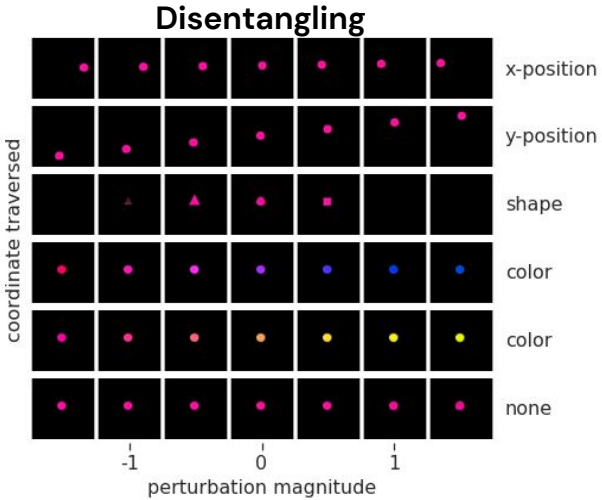
Want to learn more?

COBRA: Data-Efficient Model-Based RL through Unsupervised Object Discovery and Curiosity-Driven Exploration, Watters, Matthey et al, arxiv 2019

During unsupervised exploration stage learn:

- Object decomposition and feature disentangling
- Action-conditioned transition model of the environment
- Curiosity-driven exploration policy





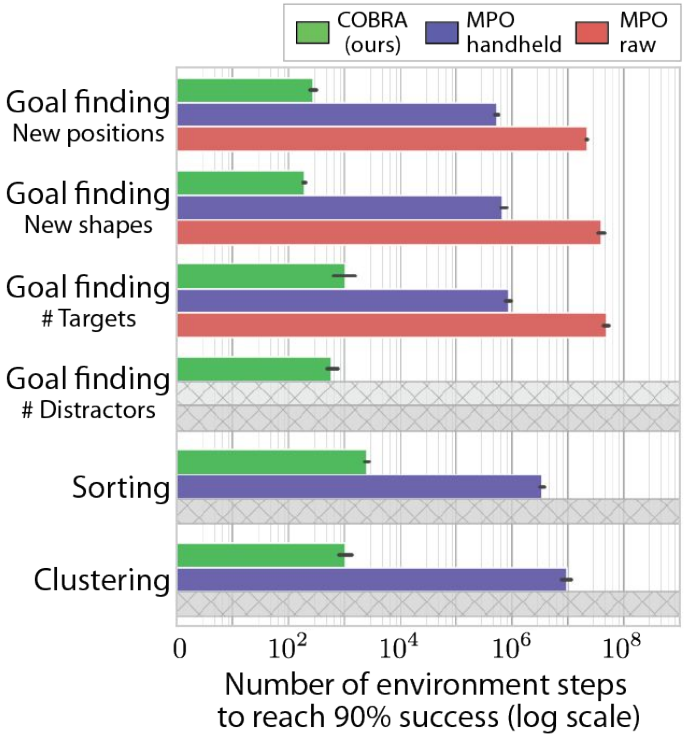
Utility of object-based representations

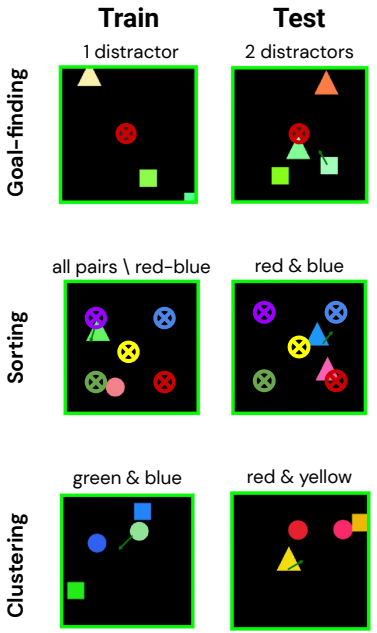
Want to learn more?

COBRA: Data-Efficient Model-Based RL through Unsupervised Object Discovery and Curiosity-Driven Exploration, Watters, Matthey et al, arxiv 2019

Task learning reduced to to learning reward function for model-based search.

- Better data efficiency
- Better generalisation



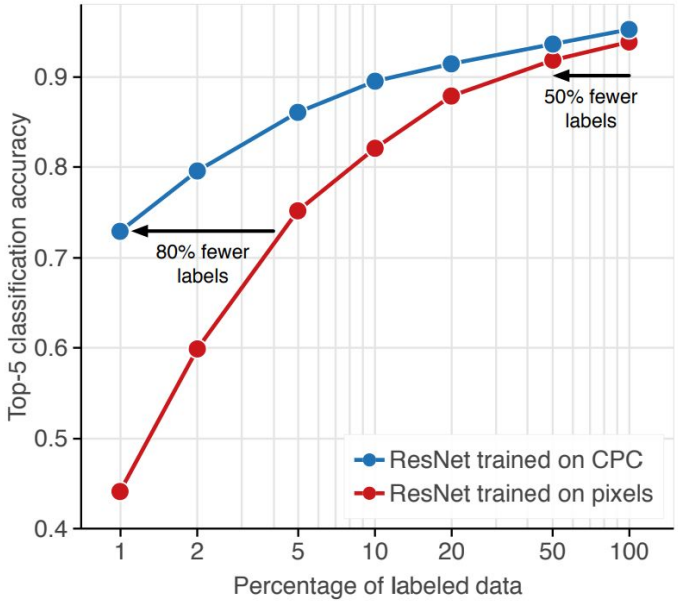


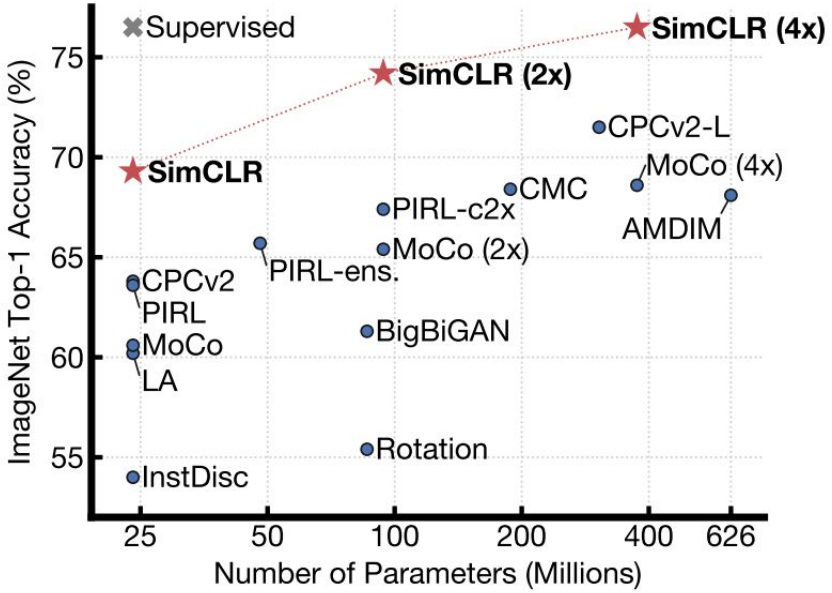
Utility of contrastive methods

Want to learn more?

Data-Efficient Image Recognition with Contrastive Predictive Coding, Olivier J. Hénaff et al, ICML 2020

A Simple Framework for Contrastive Learning of Visual Representations, Chen et al, ICML 2020



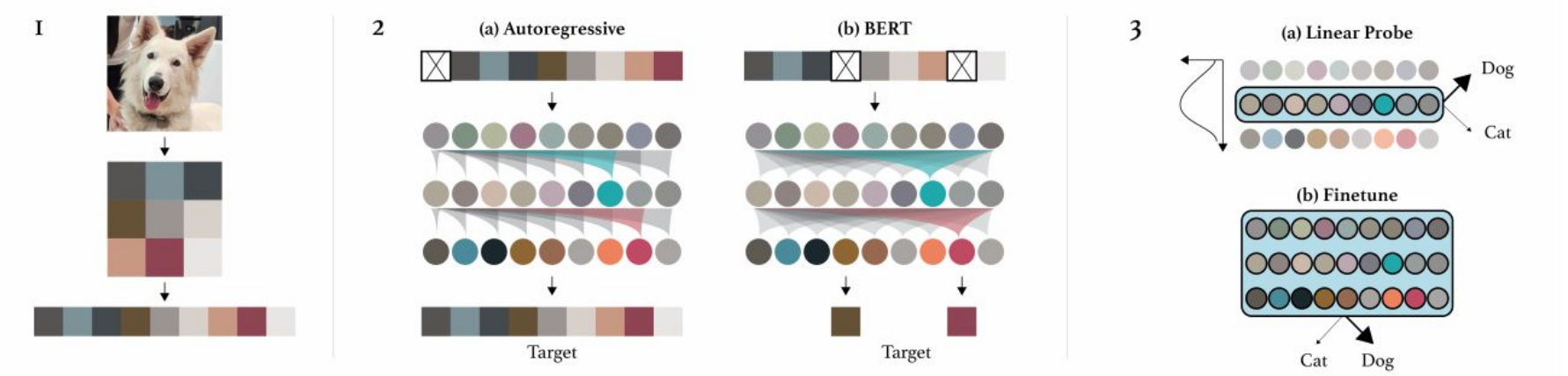


Utility of attention-based methods

Want to learn more?

Generative Pretraining from Pixels, Chen et al, ICML 2020

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

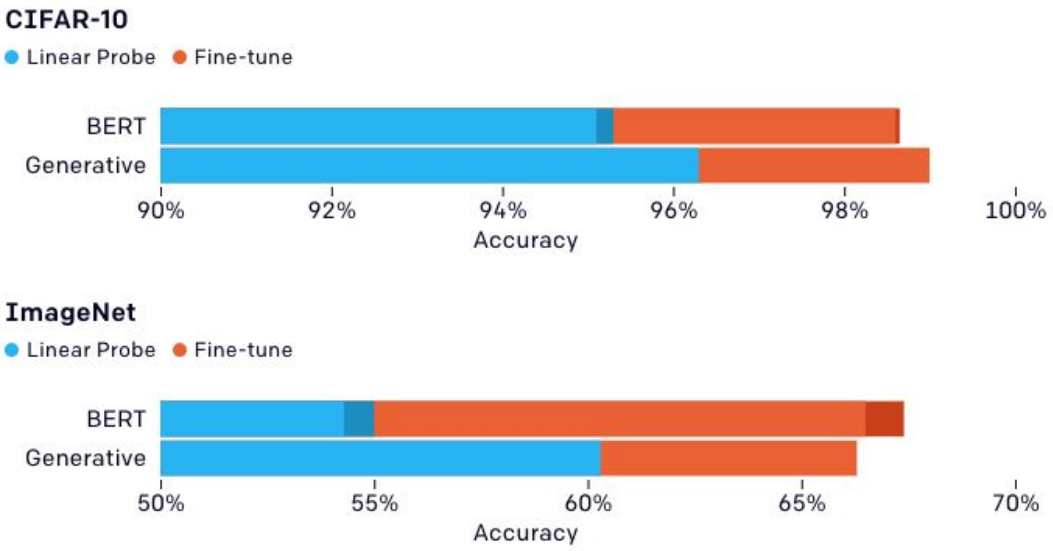


Utility of attention-based methods

Want to learn more?

Generative Pretraining from Pixels, Chen et al, ICML 2020

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



Utility of attention-based methods

Want to learn more?

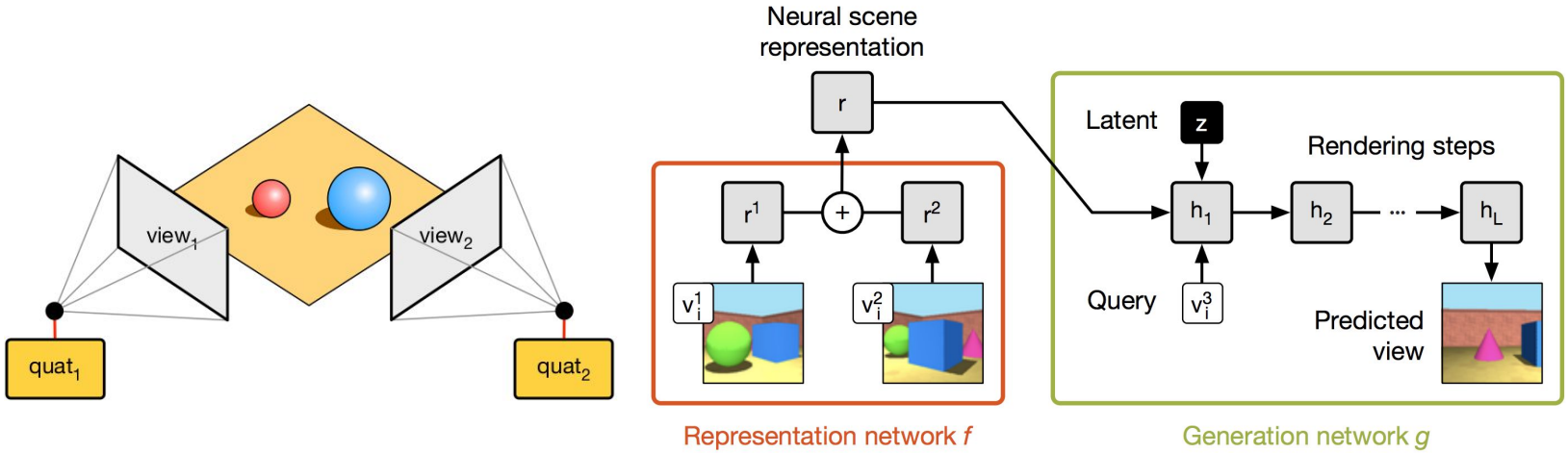
Generative Pretraining from Pixels, Chen et al, ICML 2020

EVALUATION	MODEL	ACCURACY	PRE-TRAINED ON IMAGENET W/O LABELS	PRE-TRAINED ON IMAGENET W/ LABELS
CIFAR-10 Linear Probe	ResNet-152[50]	94.0		✓
	SimCLR[12]	95.3	✓	
	iGPT-L 32x32	**96.3**	✓	
CIFAR-100 Linear Probe	ResNet-152	78.0		✓
	SimCLR	80.2	✓	
	iGPT-L 32x32	**82.8**	✓	
STL-10 Linear Probe	AMDIM-L[13]	94.2	✓	
	iGPT-L 32x32	**95.5**	✓	
CIFAR-10 Fine-tune	AutoAugment[51]	98.5		
	SimCLR	98.6	✓	
	GPipe[15]	**99.0**		✓
	iGPT-L	**99.0**	✓	
CIFAR-100 Fine-tune	iGPT-L	88.5	✓	
	SimCLR	89.0	✓	
	AutoAugment	89.3		
	EfficientNet[52]	**91.7**		✓

Utility of better belief representations (GQN)

Want to learn more?

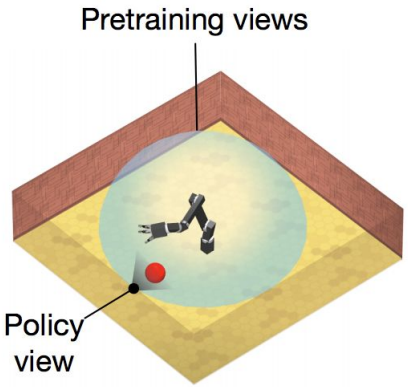
Neural Scene Representation and Rendering, Eslami et al, Science 2018

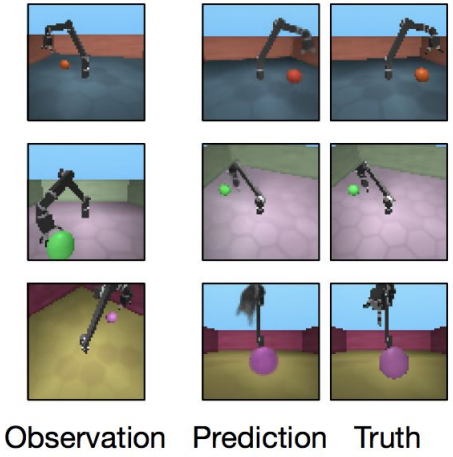


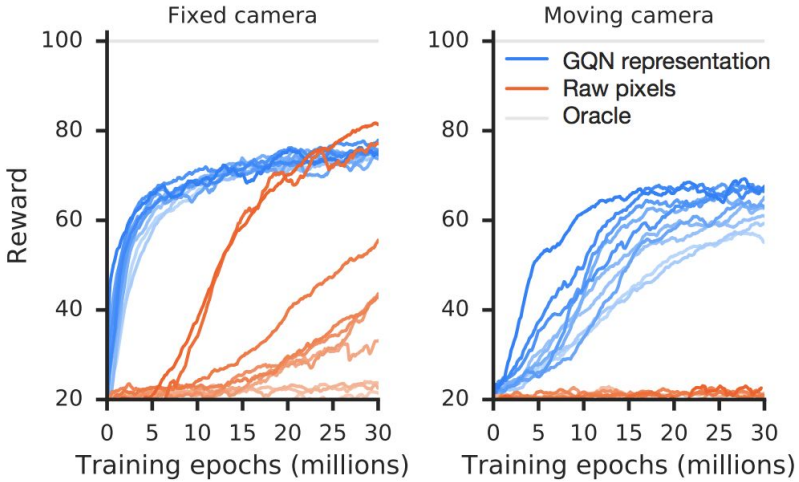
Utility of better belief representations (GQN)

Want to learn more?

Neural Scene Representation and Rendering, Eslami et al, Science 2018



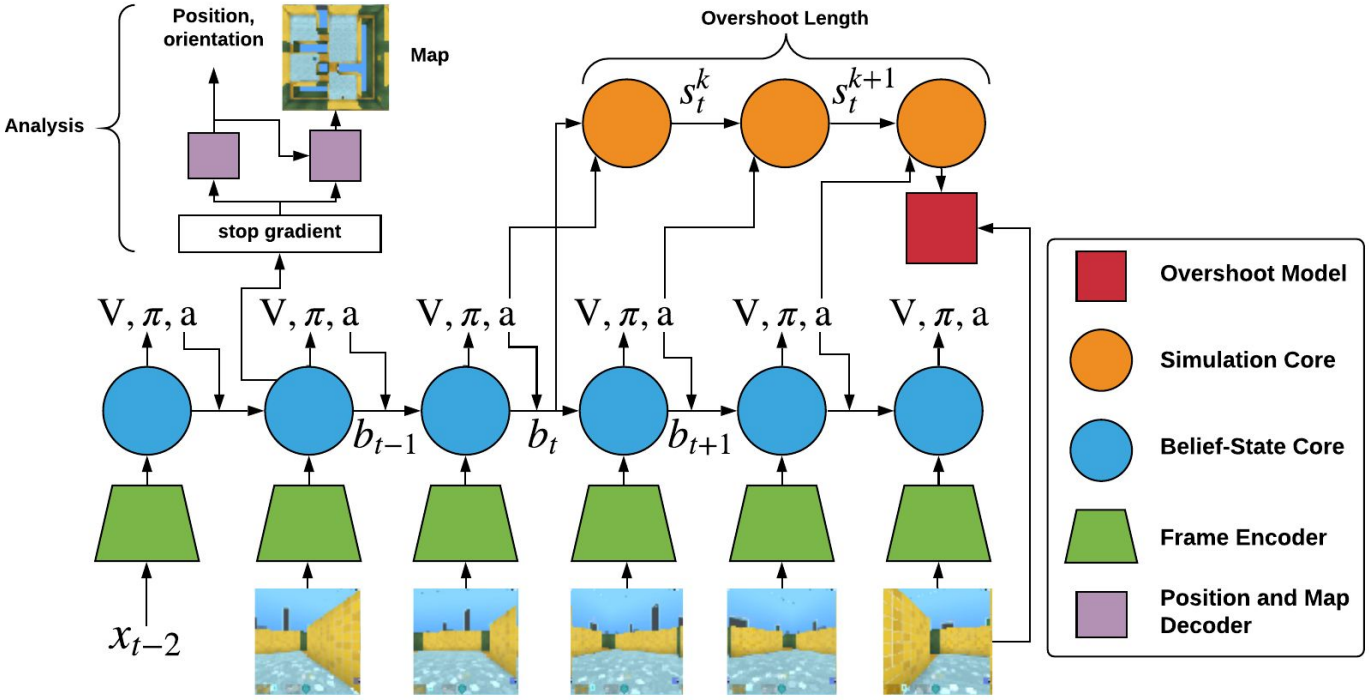




Utility of better belief representations (SimCore)

Want to learn more?

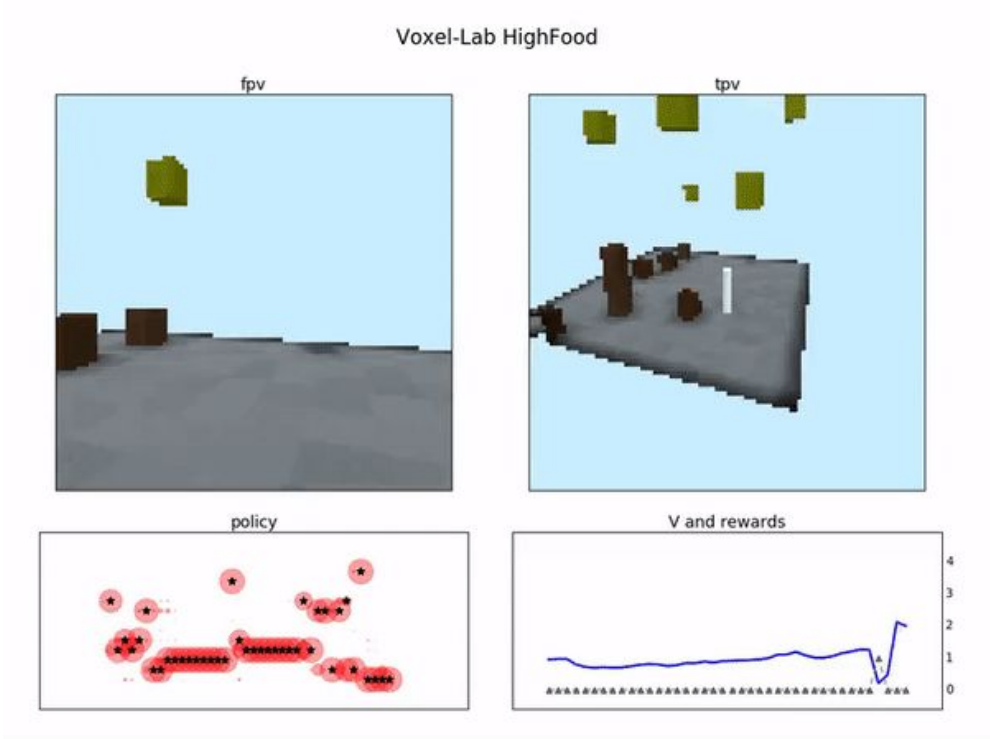
Shaping Belief States with Generative Environment Models for RL, Gregor et al, NeurIPS 2019

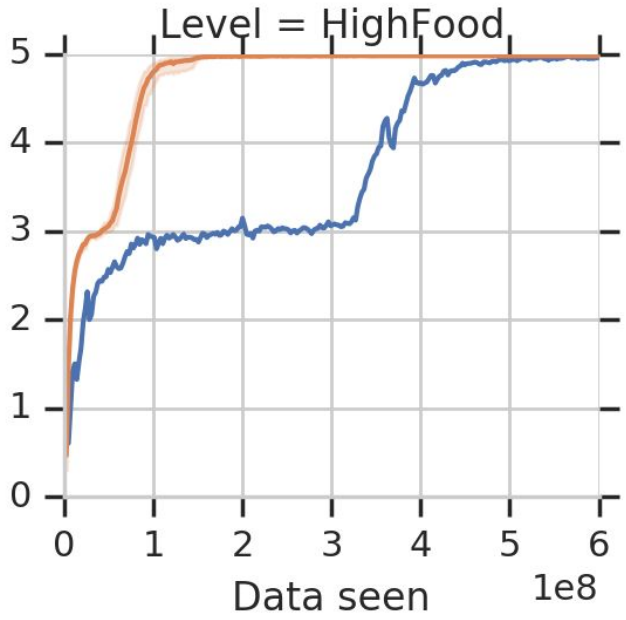


Utility of better belief representations (SimCORE)

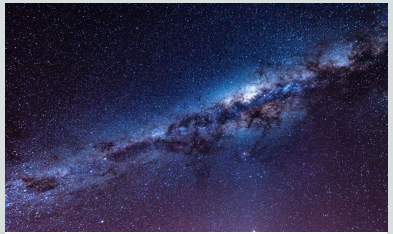
Want to learn more?

Shaping Belief States with Generative Environment Models for RL, Gregor et al, NeurIPS 2019



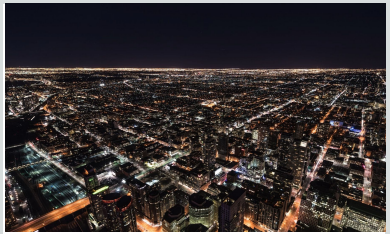


Take away messages



01

Representations are **useful abstractions** that make downstream computations and tasks more **efficient**.



02

Representation learning problem is **under-specified**.

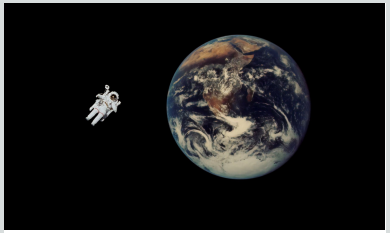
Current approaches tend to tradeoff **generality vs interpretability.**

Recent progress is surprisingly impressive.

03

The **diverse "model zoo"** can be understood using **simple theoretical taxonomy**.

Thinking about **density modelling**, **manifolds** and the level of **modelling detail** may help understand the tradeoff and areas of improvement for different methods.



04

Multi-disciplinary insights may help resolve some of the under-specification.

Different representations **help with different tasks**.

No agreed upon **evaluation method**.

video credit: Francis Vachon
www.francisvachon.com

Thank you

Questions