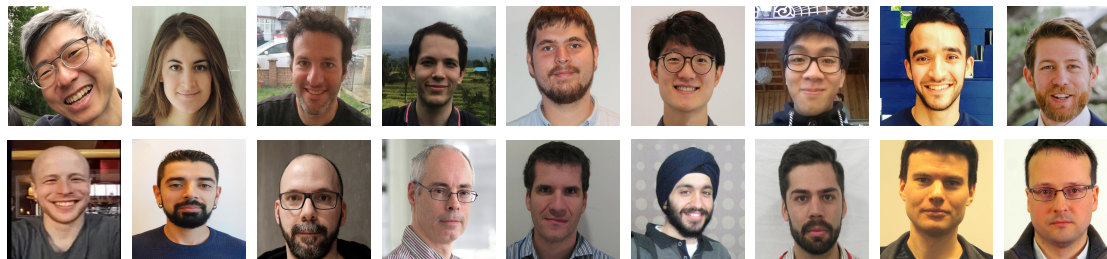


On Statistical Thinking in Deep Learning

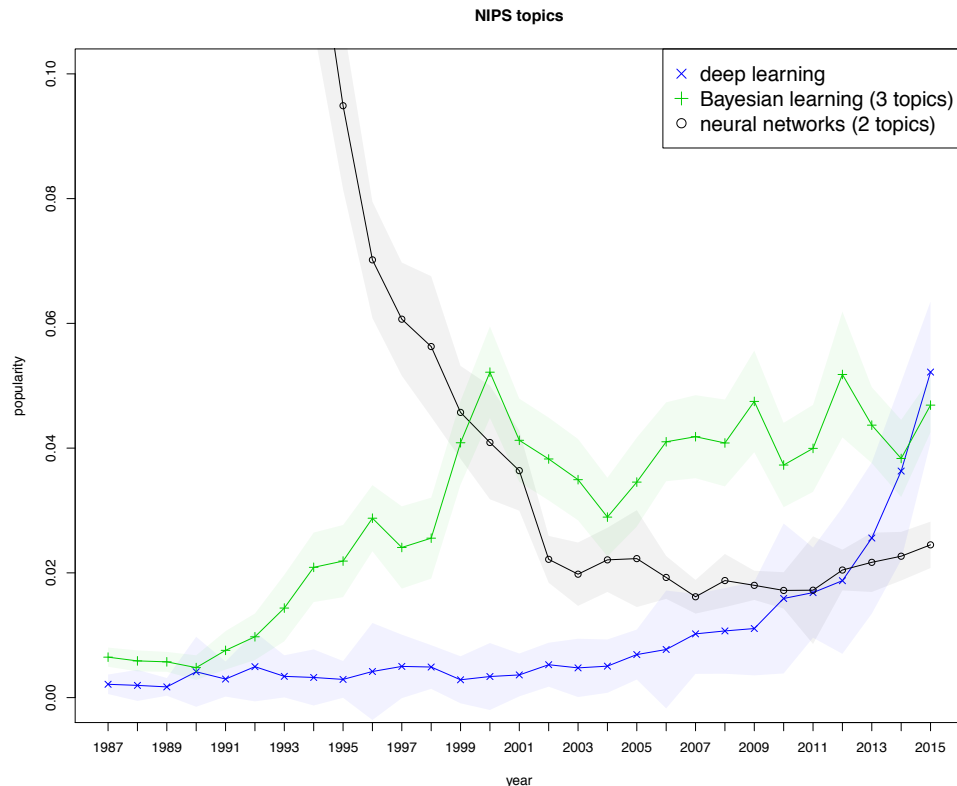
Yee Whye Teh
Statistics @ Oxford
DeepMind

<http://csml.stats.ox.ac.uk/people/teh/>



Yee Whye Teh

Probabilistic and Deep Learning @ NeurIPS '87-'15



Top words in each topic:

neural networks

architecture	propagation
recurrent	feedforward
back	feedback
activation	backpropagation
outputs	hinton
forward	connectionist
net	simulation
connections	development
connection	topology
nets	represented
connected	structures
parallel	role

deep learning

deep	rnn
layers	bengio
convolutional	train
mnist	hinton
sgd	unsupervised
rbm	boltzmann

Probabilistic learning

chain	predictive
gibbs	importance
carlo	hyperparameters
monte	particle
mcmc	stationary
sampler	proposal

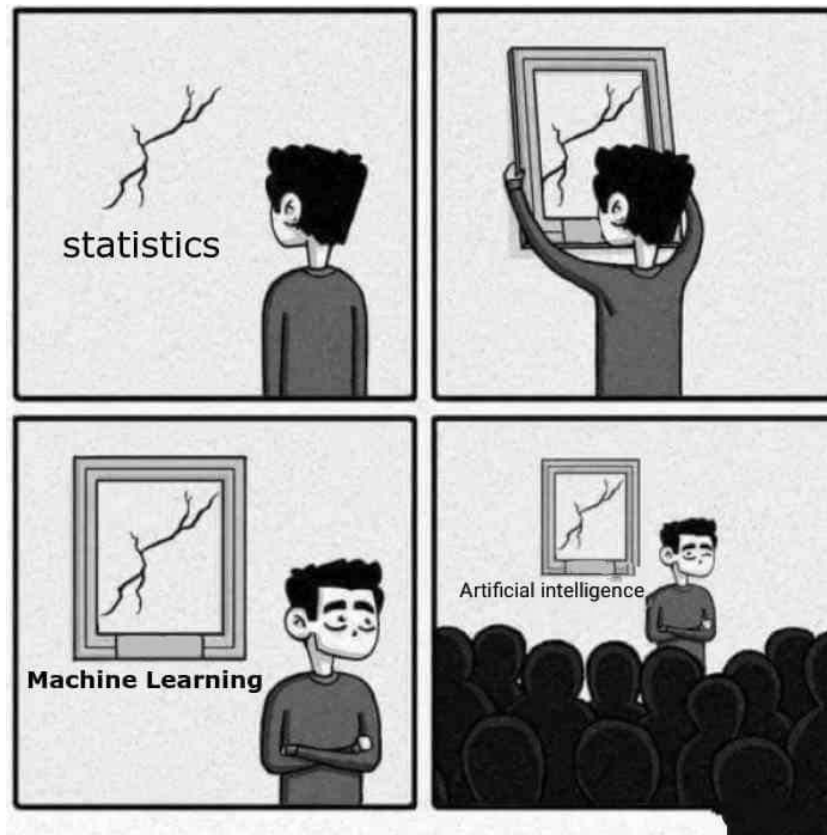
belief	messages
graphical	passing
propagation	assignment
marginal	potentials
message	pairwise
marginals	mrf

generative	bayes
priors	poisson
missing	inferred
gamma	jordan
dirichlet	counts
dependencies	multinomial

Analysis from a Bayesian nonparametric dynamic topic model
 Poisson random fields for dynamic feature models [Perrone et al 2016]



Statistics vs Machine Learning vs Artificial Intelligence



“When you’re fundraising, it’s AI.
When you’re hiring, it’s ML.
When you’re implementing, it’s
logistic regression.”

— everyone on Twitter ever



Actual Contents of Talk

- Meta-learning stochastic processes with neural processes:
 - Conditional neural processes. Garnelo et al. ICML 2018. arXiv:1807.01613
 - Neural processes. Garnelo et al. ICML 2018 Workshop on Deep Generative Models. arXiv:1807.01622
 - Attentive neural processes. Kim et al. ICLR 2019. arXiv:1901.05761
 - Empirical evaluation of neural process objectives. Le et al. NeurIPS 2018 Workshop on Bayesian Deep Learning.
 - Meta-learning surrogate models for sequential decision making. Galashov et al. arXiv:1903.11907
- Probabilistic symmetries and invariant neural networks. Bloem-Reddy and Teh. arXiv:1901.06082



Few Shot Image Classification

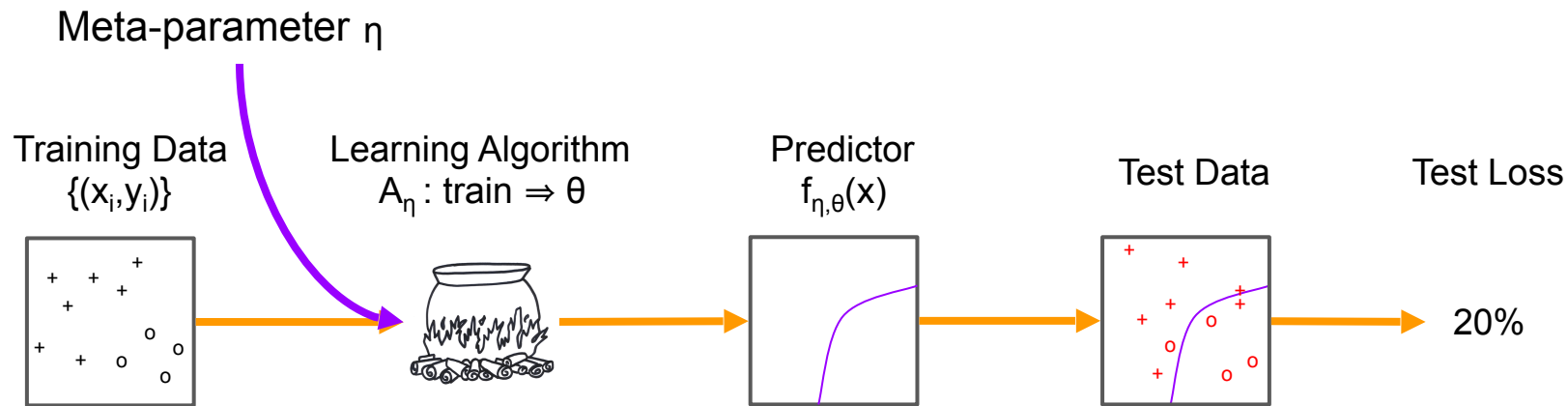


Few Shot Image Classification

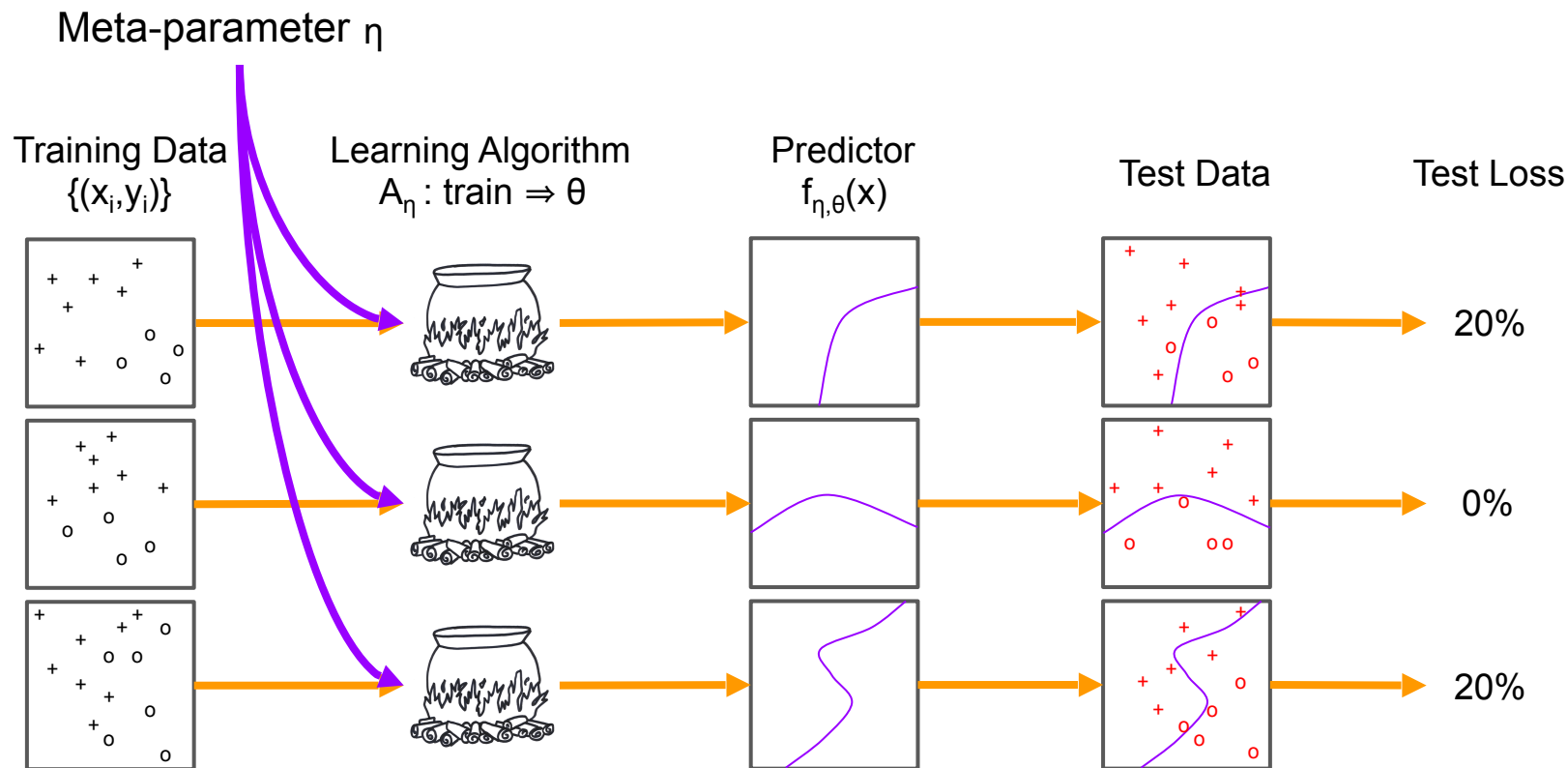
Paper (only latest shown)	<i>minilImageNet test accuracy</i>	
	5-way 1-shot	5-way 5-shot
Gidaris & Komodakis (2018)	56.20 $\pm$ 0.86%	73.00 $\pm$ 0.64%
Bauer & Rojas-Carulla (2017)	56.30 $\pm$ 0.40%	73.90 $\pm$ 0.30%
Oreshkin et al., (2018)	59.50 $\pm$ 0.30%	76.70 $\pm$ 0.30%
Siyuan Qiao et al. (2017)	59.60 $\pm$ 0.41%	73.74 $\pm$ 0.19%
LEO (Rusu et al 2018)	60.06 $\pm$ 0.05%	75.72 $\pm$ 0.18%



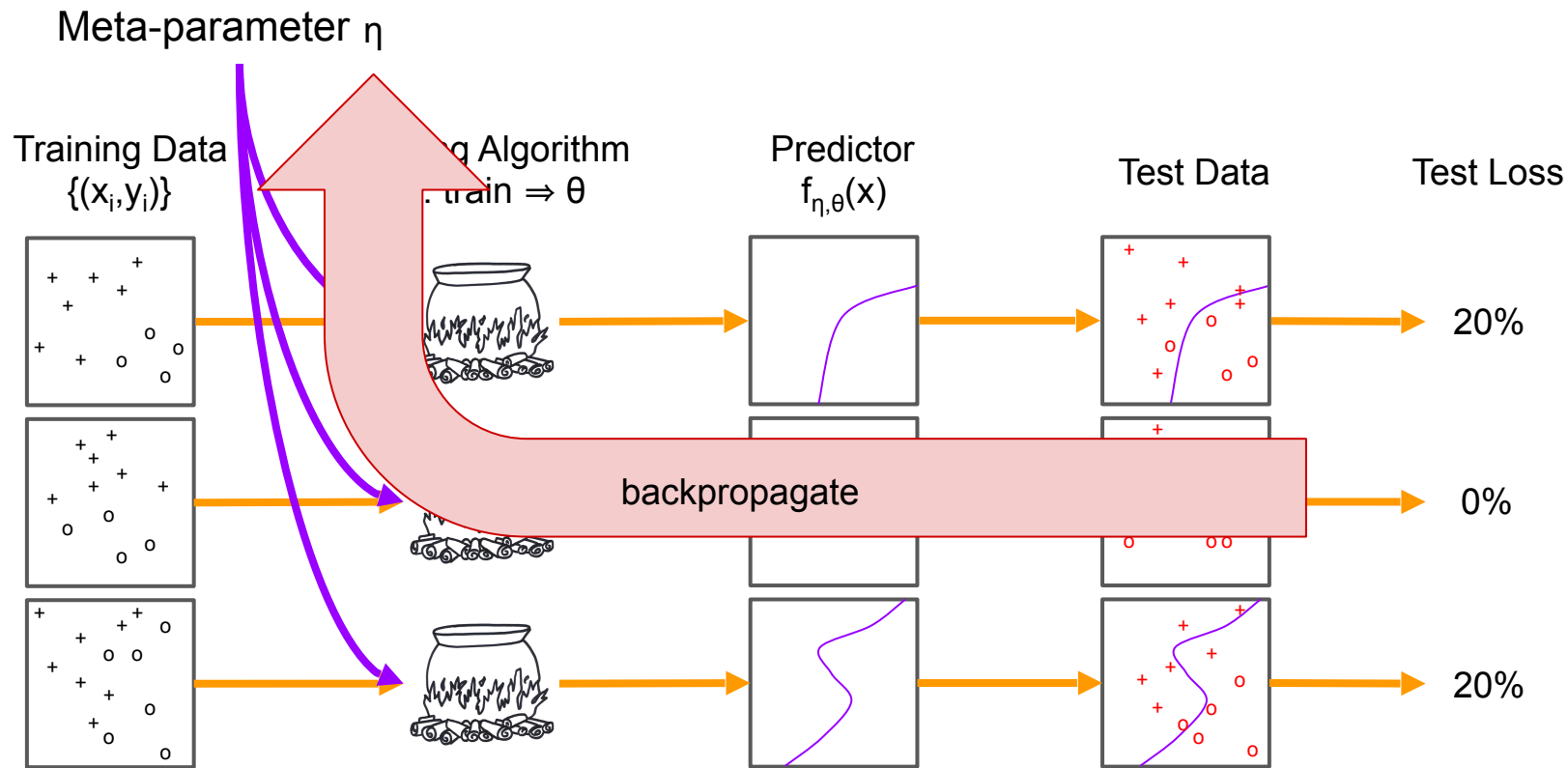
Optimisation Perspective on Meta-Learning



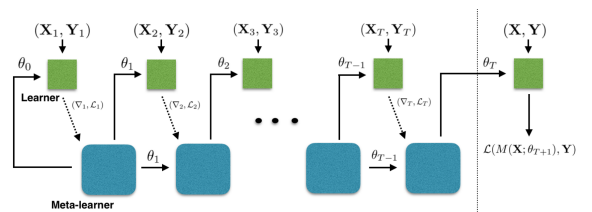
Optimisation Perspective on Meta-Learning



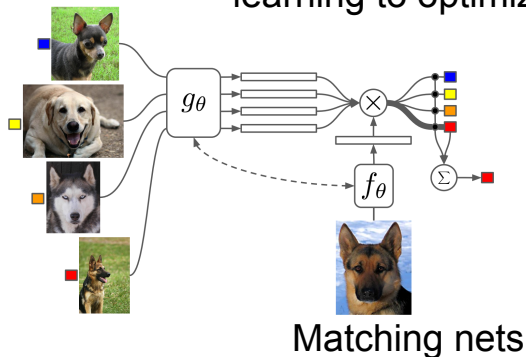
Optimisation Perspective on Meta-Learning



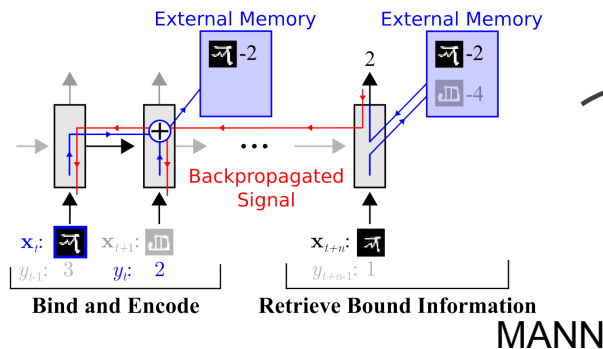
Many (Supervised) Meta-Learning Methods



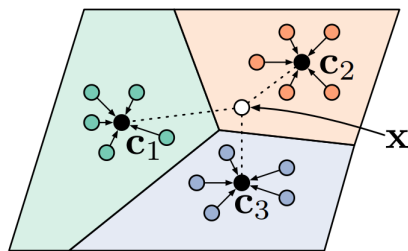
LSTM meta-learner,
learning to optimize



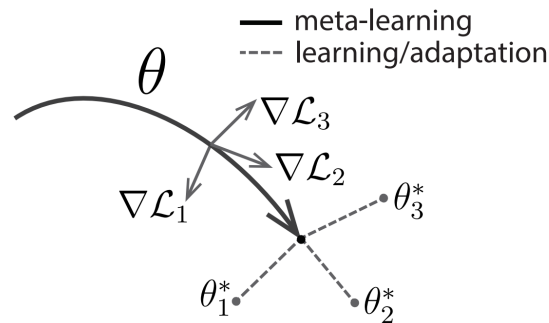
Matching nets



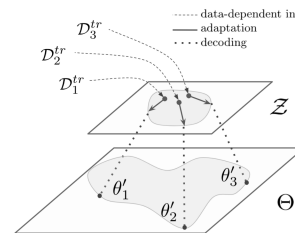
MANN



Prototypical nets



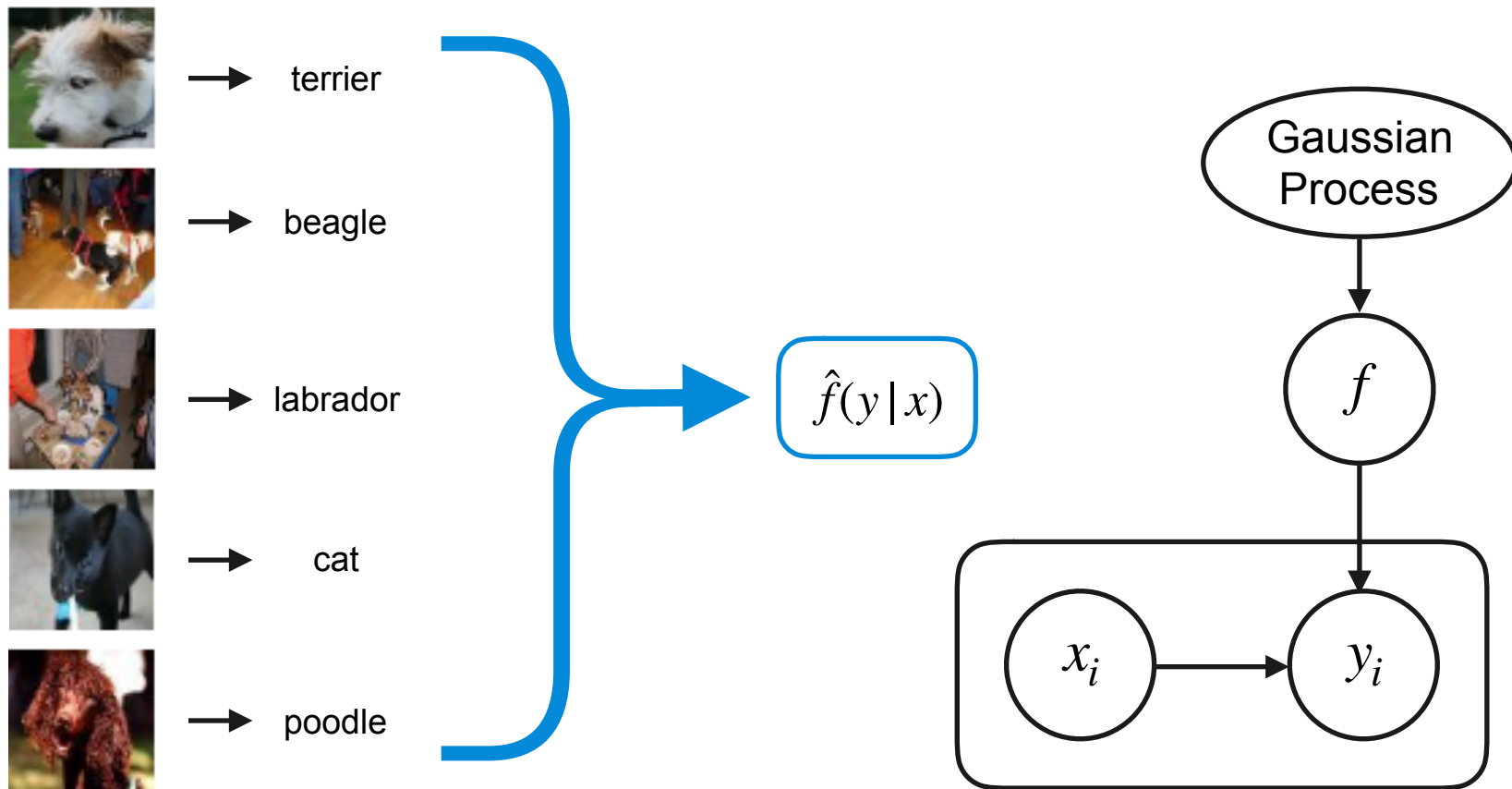
MAML



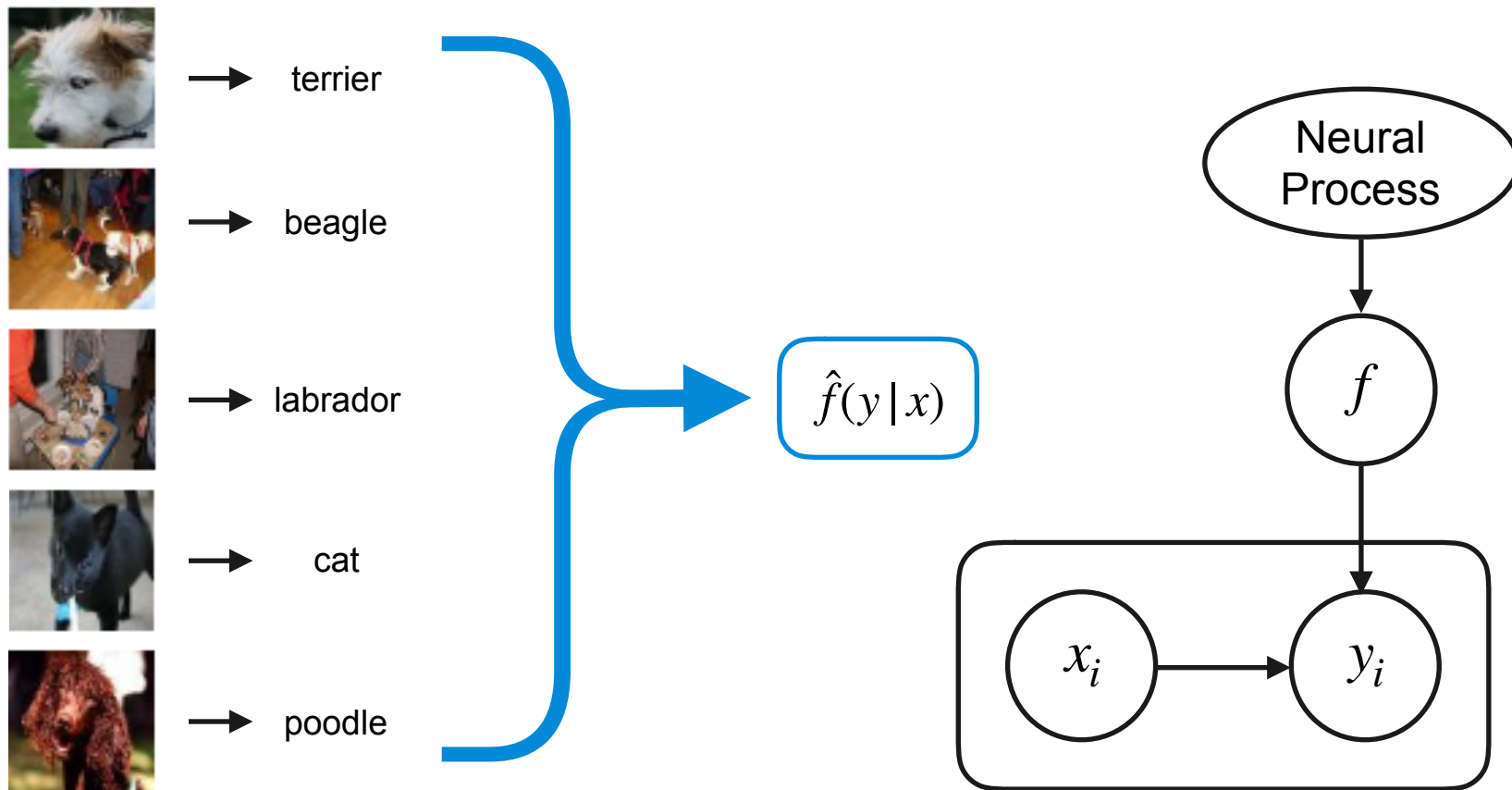
LEO



Probabilistic Perspective on Meta-Learning



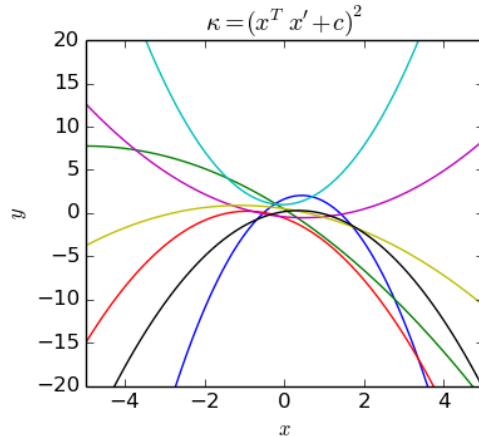
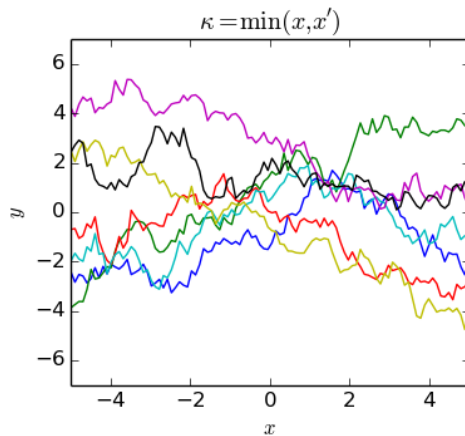
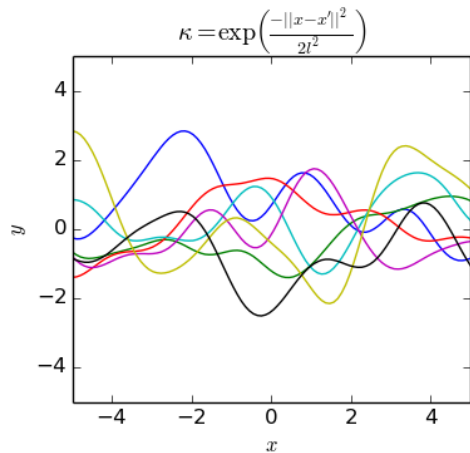
Probabilistic Perspective on Meta-Learning



Specifying Stochastic Processes

- Gaussian processes are typically described via marginal distributions:

$$\begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_t) \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \mu(x_1) \\ \mu(x_2) \\ \vdots \\ \mu(x_t) \end{pmatrix}, \begin{pmatrix} K(x_1, x_1) & K(x_1, x_2) & \cdots & K(x_1, x_t) \\ K(x_2, x_1) & K(x_2, x_2) & \cdots & K(x_2, x_t) \\ \vdots & \vdots & \ddots & \vdots \\ K(x_t, x_1) & K(x_t, x_2) & \cdots & K(x_t, x_t) \end{pmatrix} \right)$$



Specifying Stochastic Processes

- Gaussian processes can equivalently be described via its conditional distributions:

$$f(x_{t+1})|f(x_1) = y_1, \dots, f(x_t) = y_t \\ \sim \mathcal{N}(\mu(x_{t+1}) + K_{t+1,1:t}K_{1:t,1:t}^{-1}y_{1:t}, K_{t+1,t+1} - K_{t+1,1:t}K_{1:t,1:t}^{-1}K_{1:t,t+1})$$

- In general, stochastic processes can also be described using a consistent family of conditional distributions:

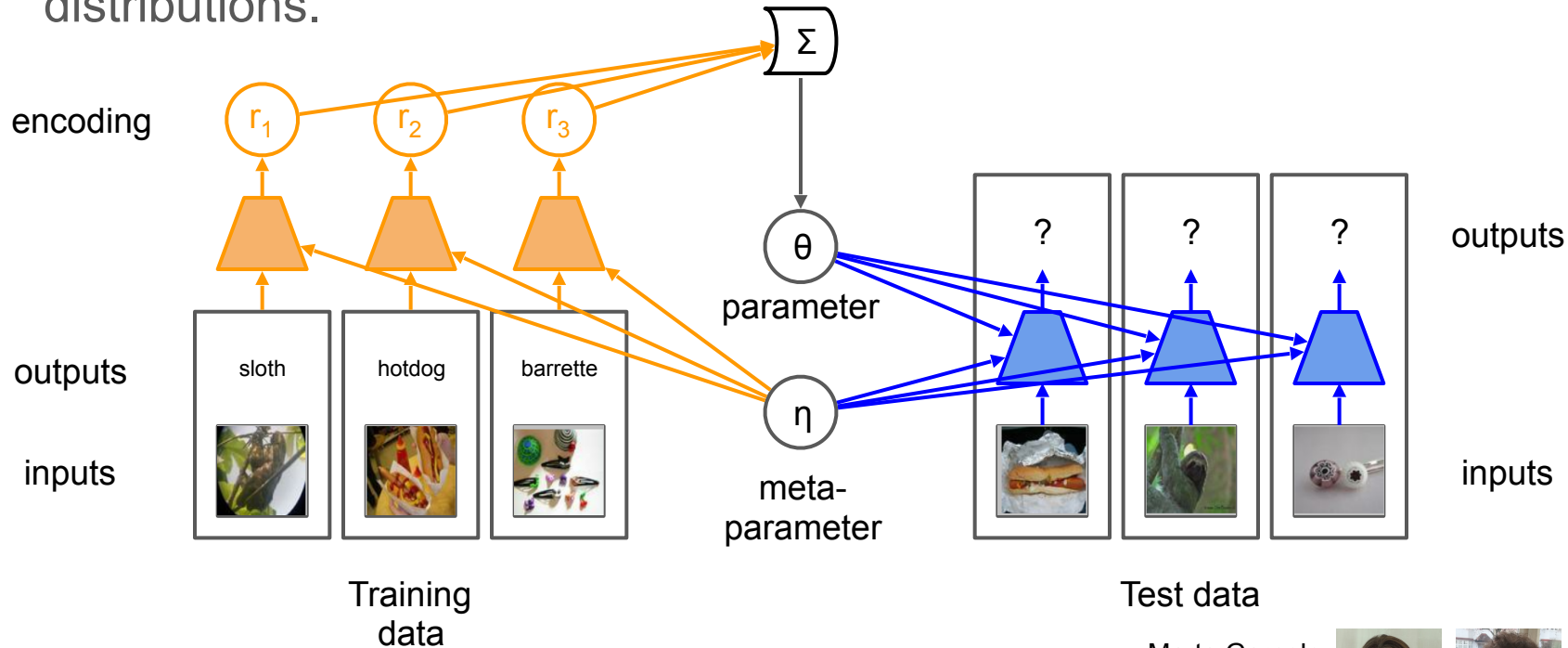
$$\mathbb{P}(f(x_{t+1}) = y_{t+1}|f(x_1) = y_1, \dots, f(x_t) = y_t)$$

for training sets $\{\mathbf{x}_{1:t}, \mathbf{y}_{1:t}\}$ and test sets $\{\mathbf{x}_{t+1}, \mathbf{y}_{t+1}\}$.

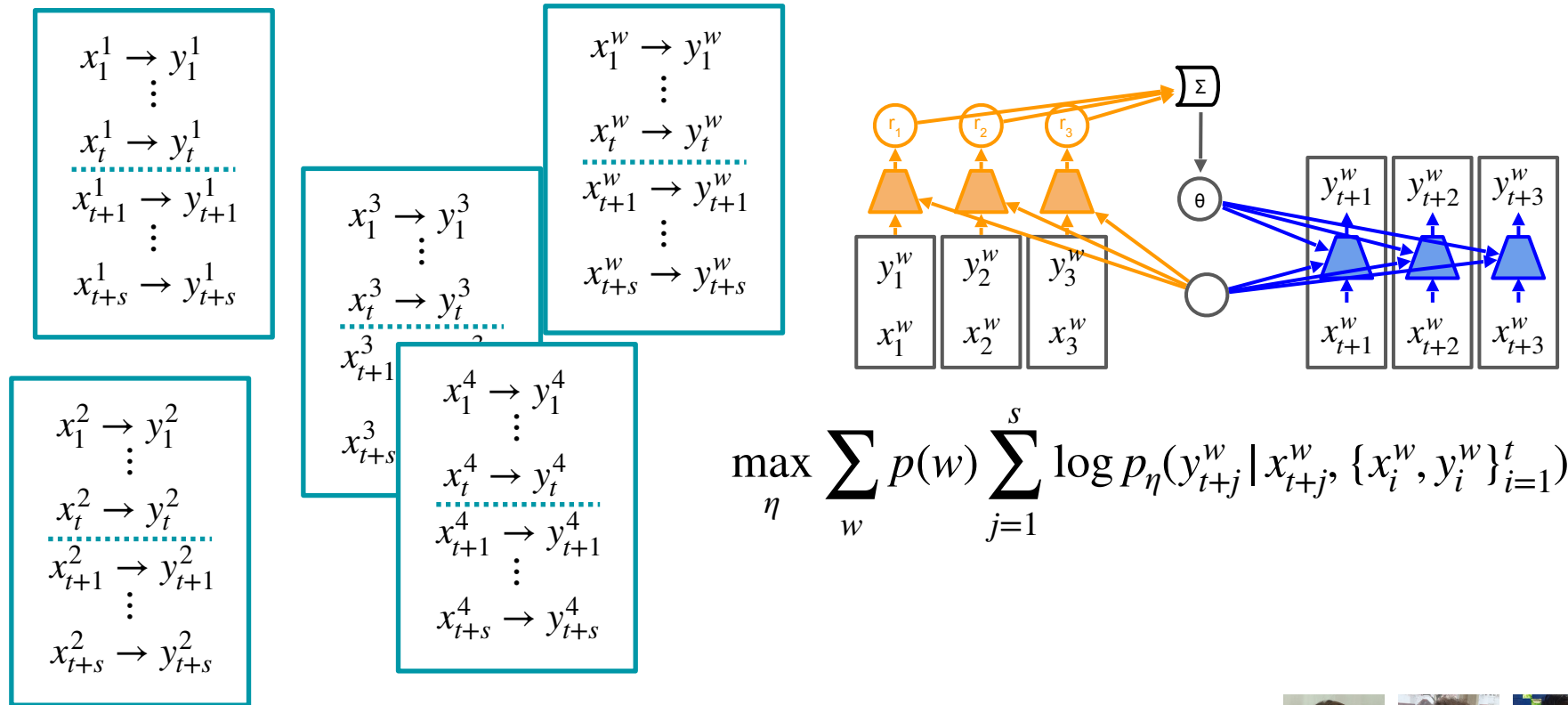


Neural Processes: Learning Neural Stochastic Processes

- Use neural networks to parameterise and learn the conditional distributions.



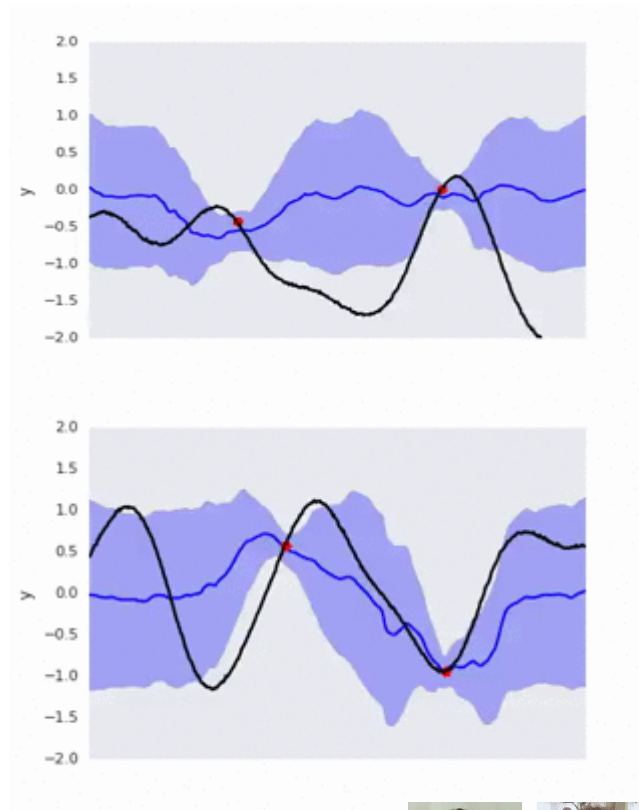
Neural Processes: Learning Neural Stochastic Processes



Neural Processes

Task = Function on 1D space.

Given training points, use neural processes to predict mean and std of function values at other locations.



Yee Whye Teh

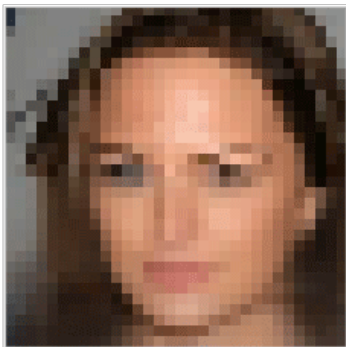
Marta Garnelo
Dan Rosenbaum
Ali Eslami



Image Super-resolution

Task = Image = Function on 2D space.

Bottom half prediction



Super-resolution

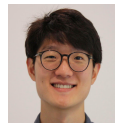
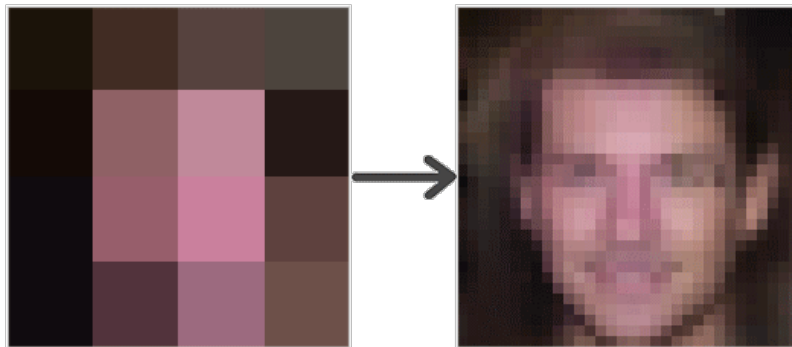
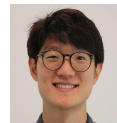
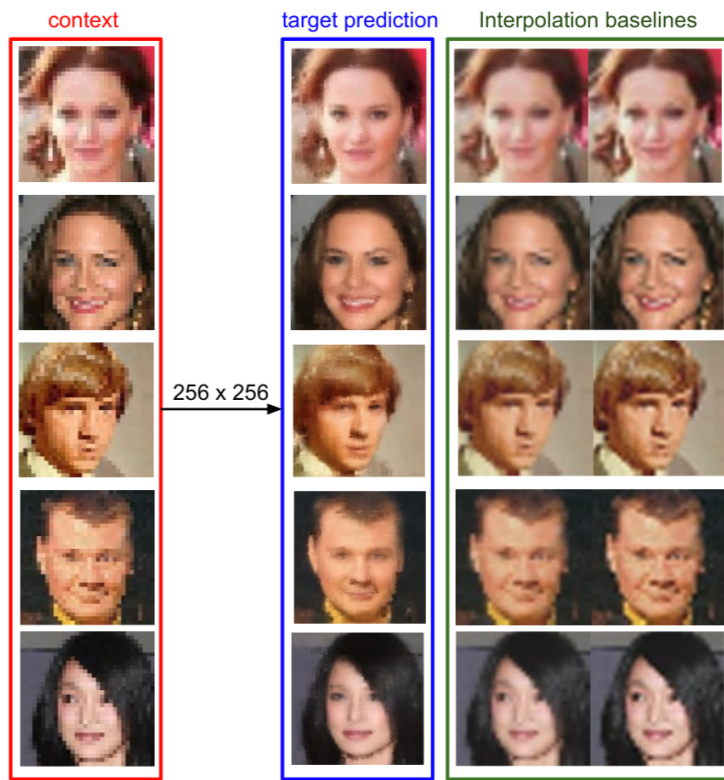
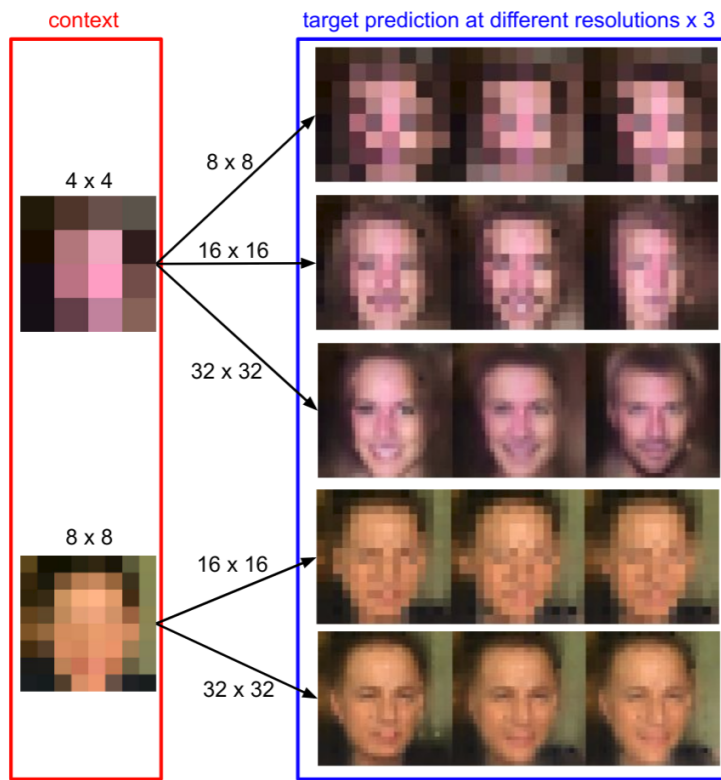
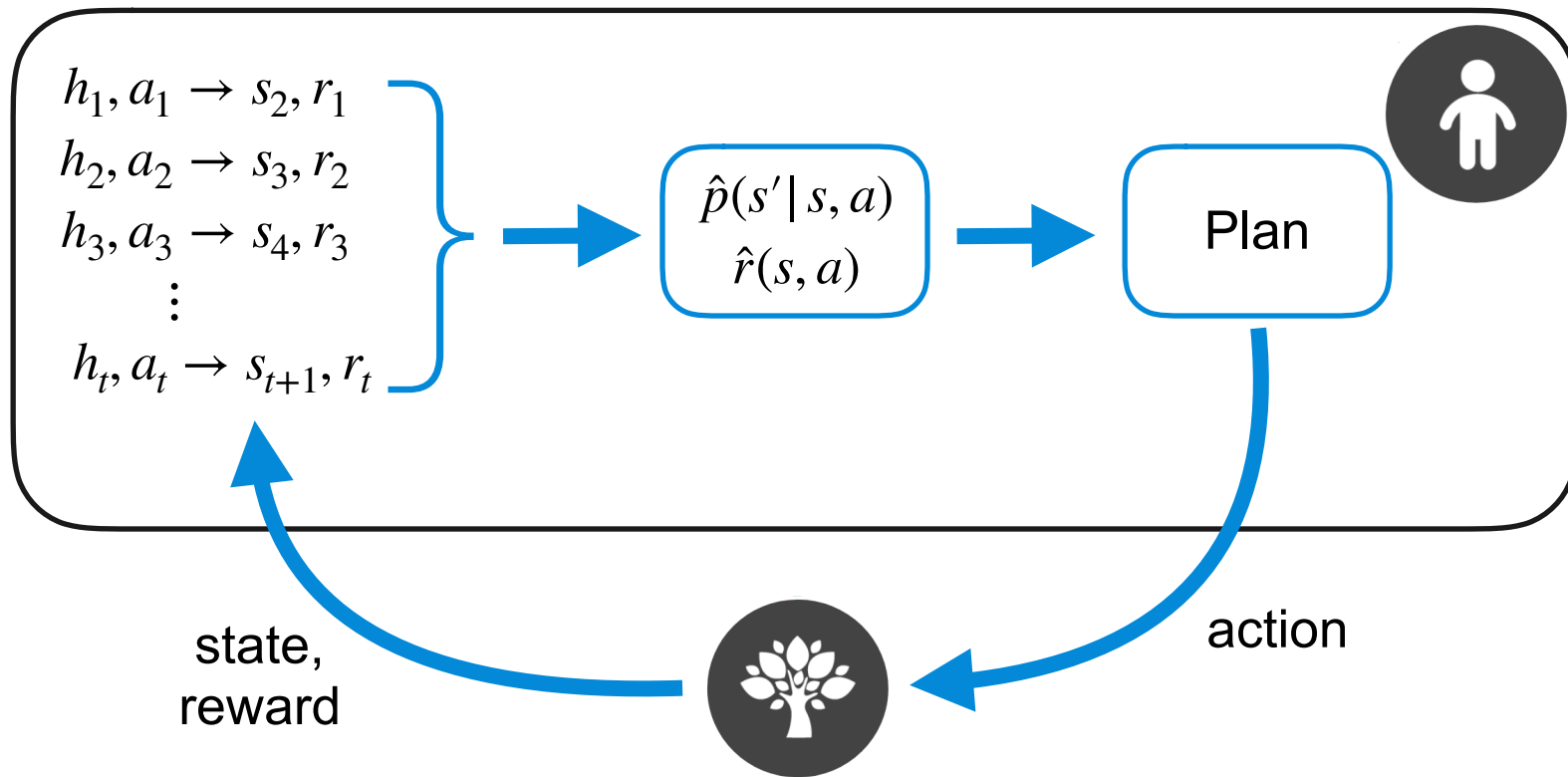


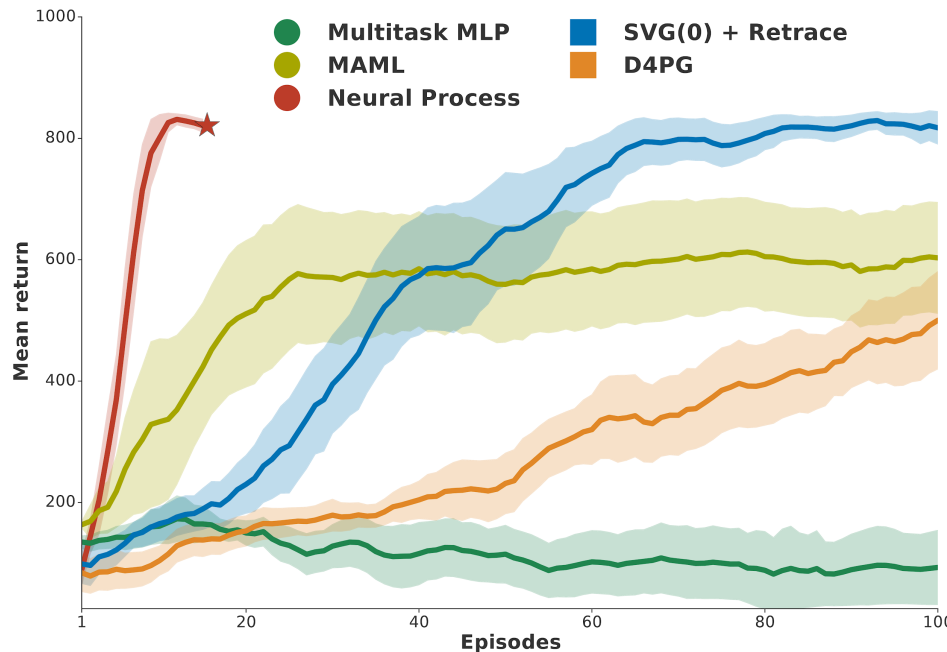
Image Super-resolution



Efficient Model-based Reinforcement Learning



Cart Pole

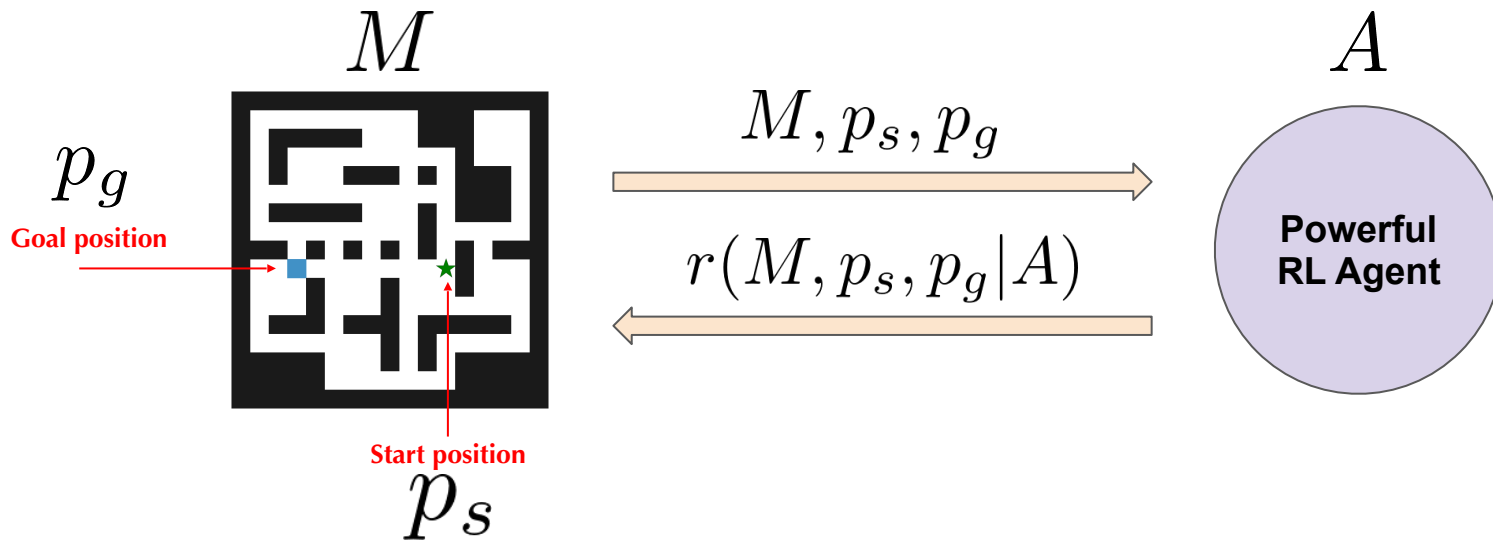


Yee Whye Teh

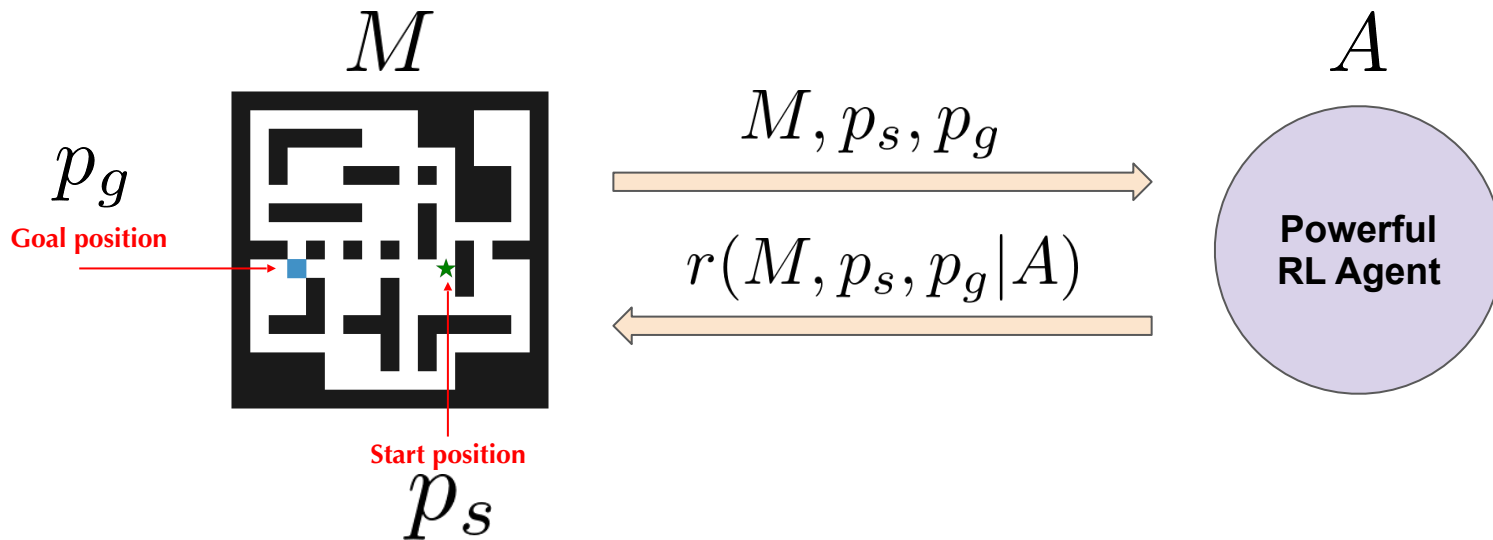
Jonathan Schwarz
Alex Galashov



Adversarial Testing of RL Agents



Adversarial Testing of RL Agents



Bayesian
Optimization

$$\min_{M, p_s, p_g} r(M, p_s, p_g | A)$$

$(M, p_s, p_g, A) \sim p(\mathcal{T})$ - training & holdout samples (agents, mazes, positions)

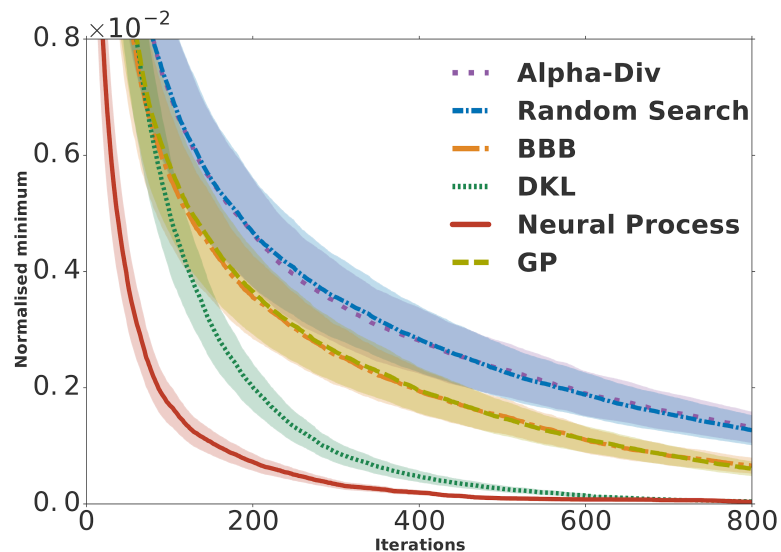


Yee Whye Teh

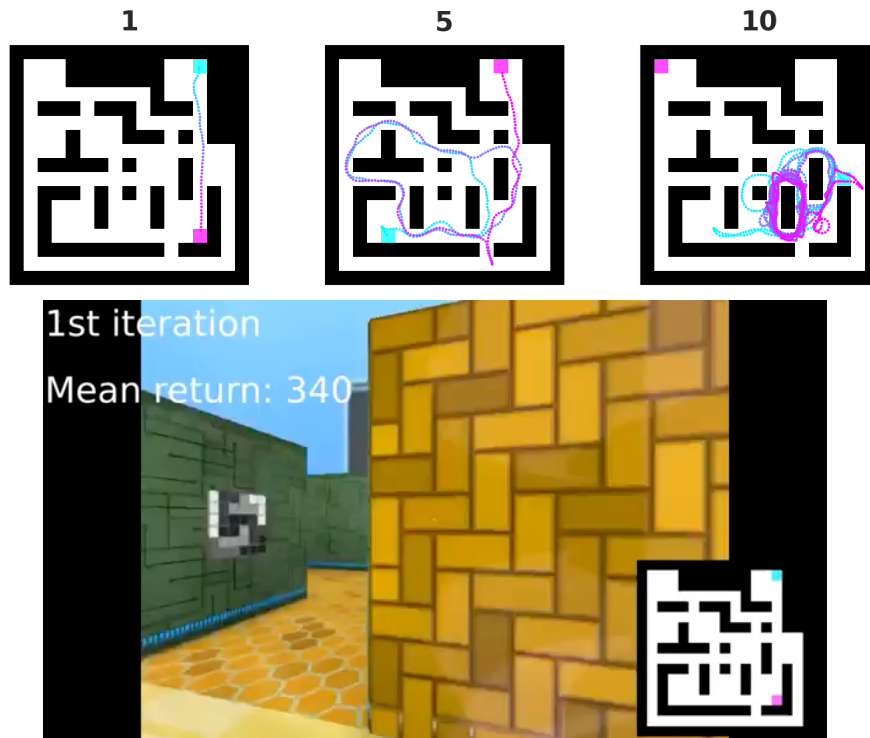
Jonathan Schwarz
Alex Galashov



Adversarial Testing of RL Agents



Bayesian optimisation iterations

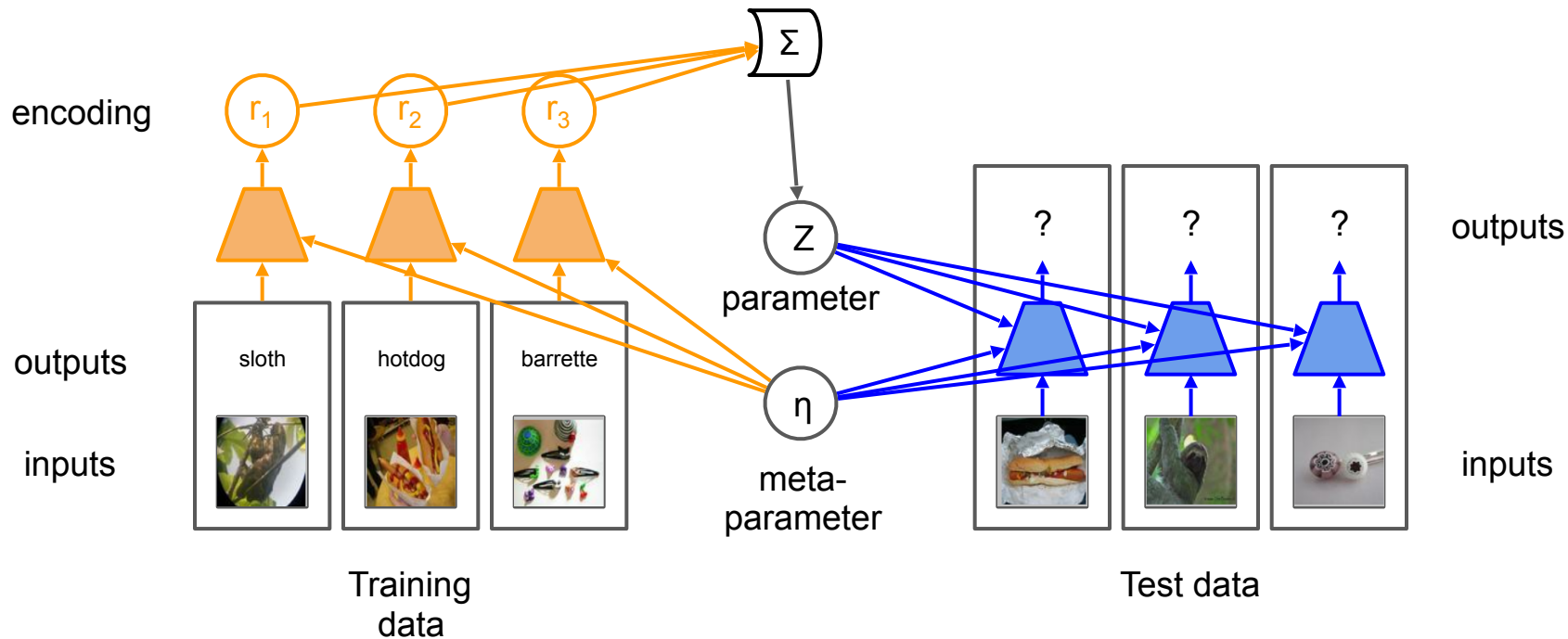


Yee Whye Teh

Jonathan Schwarz
Alex Galashov



Permutation-Invariance in Neural Processes



Actual Contents of Talk

- Meta-learning stochastic processes with neural processes:
 - Conditional neural processes. Garnelo et al. ICML 2018. arXiv:1807.01613
 - Neural processes. Garnelo et al. ICML 2018 Workshop on Deep Generative Models. arXiv:1807.01622
 - Attentive neural processes. Kim et al. ICLR 2019. arXiv:1901.05761
 - Empirical evaluation of neural process objectives. Le et al. NeurIPS 2018 Workshop on Bayesian Deep Learning.
 - Meta-learning surrogate models for sequential decision making. Galashov et al. arXiv:1903.11907
- Probabilistic symmetries and invariant neural networks. Bloem-Reddy and Teh. arXiv:1901.06082



Characterising Permutation-Invariant Functions

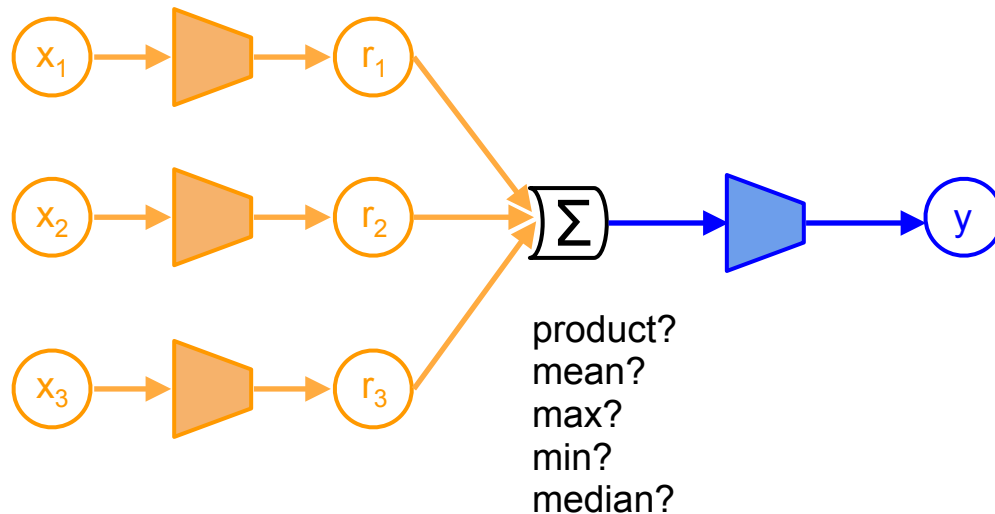
- Function $h : \mathcal{X}^n \rightarrow \mathcal{Y}$ is permutation-invariant,

$$h(\pi \cdot (x_1, \dots, x_n)) = h(x_{\pi(1)}, \dots, x_{\pi(n)}) = h(x_1, \dots, x_n)$$

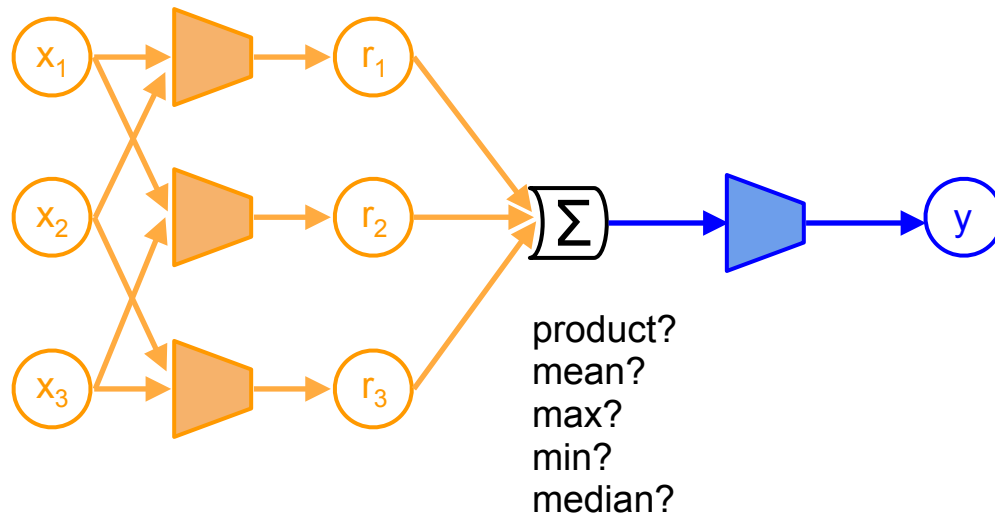
- Can we characterise the class of permutation-invariant functions?
- If we use neural networks to parameterise permutation-invariant functions, how should we choose the architecture?
- Given an architecture choice, can the neural network approximate well any arbitrary permutation-invariant function?



Characterising Permutation-Invariant Functions



Characterising Permutation-Invariant Functions



Functional Symmetry Properties

- Function $h : \mathcal{X}^n \rightarrow \mathcal{Y}^n$ is permutation-equivariant,

$$h(x_1, \dots, x_n) = (y_1, \dots, y_n)$$

$$h(\pi \cdot (x_1, \dots, x_n)) = \pi \cdot (y_1, \dots, y_n) = \pi \cdot h(x_1, \dots, x_n)$$

- Group G acting on input space $\mathcal{X}$ and output space $\mathcal{Y}$.

- G -invariant:

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

- G -equivariant:

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



Probabilistic Symmetries

- A distribution P for a random sequence $\mathbf{X}_n = (X_1, \dots, X_n)$ is exchangeable if

$$P(X_1, \dots, X_n) = P(\pi \cdot (X_1, \dots, X_n))$$

$$P(X_1 \in B_1, \dots, X_n \in B_n) = P(X_{\pi(1)} \in B_1, \dots, X_{\pi(n)} \in B_n))$$

- Exchangeability is permutation-invariance of P .
- $\mathbf{X}_{\mathbb{N}}$ is infinitely exchangeable if all length n prefixes are exchangeable.
- de Finetti's Theorem:

$\mathbf{X}_{\mathbb{N}}$ is infinitely exchangeable $\Leftrightarrow X_i | \mathcal{Q} \sim_{iid} \mathcal{Q}$ for some random $\mathcal{Q}$.



Probabilistic Symmetries for Conditional Distributions

- A conditional distribution $P(Y | X)$ is a stochastic relaxation for a function $Y = h(X)$.

- $P(Y | X)$ is G -invariant if:

$$P(Y | X) = P(Y | g \cdot X)$$

$$P(Y \in B | X \in A) = P(Y \in B | g \cdot X \in A)$$

- $P(Y | X)$ is G -equivariant if:

$$P(Y | X) = P(g \cdot Y | g \cdot X)$$

- Can we characterise the class of permutation-invariant conditional distributions?



Empirical Measure

- de Finetti's Theorem may fail for finitely exchangeable sequences.

- The empirical measure of $\mathbf{X}_n$ is

$$\mathbb{M}_{\mathbf{X}_n}(\cdot) = \sum_{i=1}^n \delta_{X_i}(\cdot)$$

- The empirical measure is a sufficient statistic: P is exchangeable iff

$$P(\mathbf{X}_n \in \cdot \mid \mathbb{M}_{\mathbf{X}_n} = m) = \mathbb{U}_m(\cdot)$$

where $\mathbb{U}_m$ is the uniform distribution over all sequences $(x_1, \dots, x_n)$ with empirical measure m .



Noise Outsourcing

- If X and Y are random variables in “nice” (e.g. Borel) spaces $\mathcal{X}$ and $\mathcal{Y}$, then there are a random variable $\eta \sim U[0,1]$ with $\eta \perp\!\!\!\perp X$ and a function $h : [0,1] \times \mathcal{X} \mapsto \mathcal{Y}$ such that

$$(X, Y) =_{a.s.} (X, h(\eta, X))$$

- Furthermore, if there is statistic $S(X)$ with $X \perp\!\!\!\perp Y \mid S(X)$, then

$$(X, Y) =_{a.s.} (X, h(\eta, S(X)))$$



Probabilistic Permutation-Invariance

- Now suppose we have random variables $\mathbf{X}_n$ and Y .
 - Y is conditionally permutation-invariant given $\mathbf{X}_n$.
 - $\mathbf{X}_n$ is marginally permutation-invariant (exchangeable).
- The empirical measure is a sufficient statistic for $\mathbf{X}_n$.
- It is also an adequate statistic for Y given $\mathbf{X}_n$:

$$P(Y | \mathbf{X}_n = \mathbf{x}_n) = P(Y | \mathbb{M}_{\mathbf{X}_n} = \mathbb{M}_{\mathbf{x}_n})$$

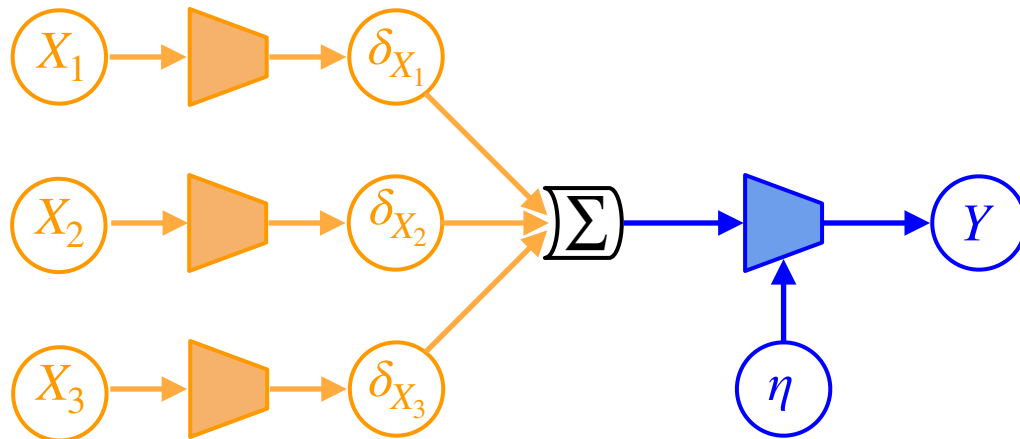
We have the conditional independence $\mathbf{X}_n \perp\!\!\!\perp Y | \mathbb{M}_{\mathbf{X}_n}$.

- Noise outsourcing...

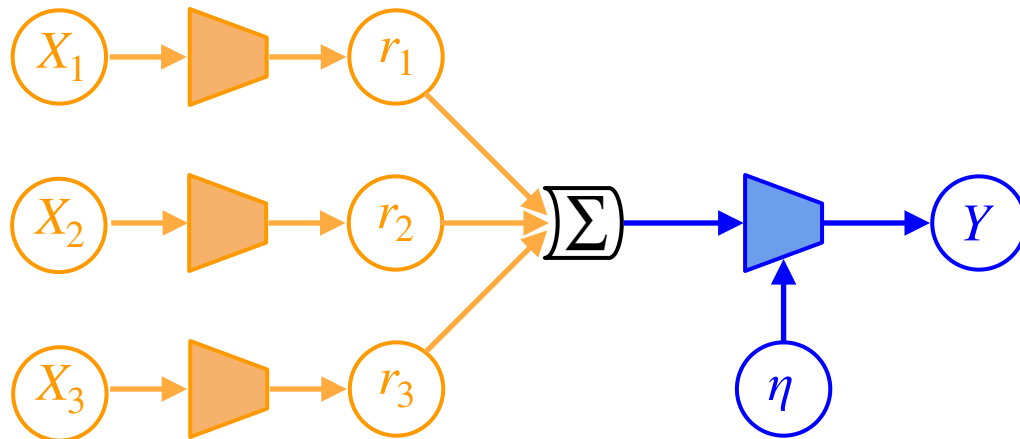
$$(\mathbf{X}_n, Y) =_{a.s.} (\mathbf{X}_n, h(\eta, \mathbb{M}_{\mathbf{X}_n}))$$



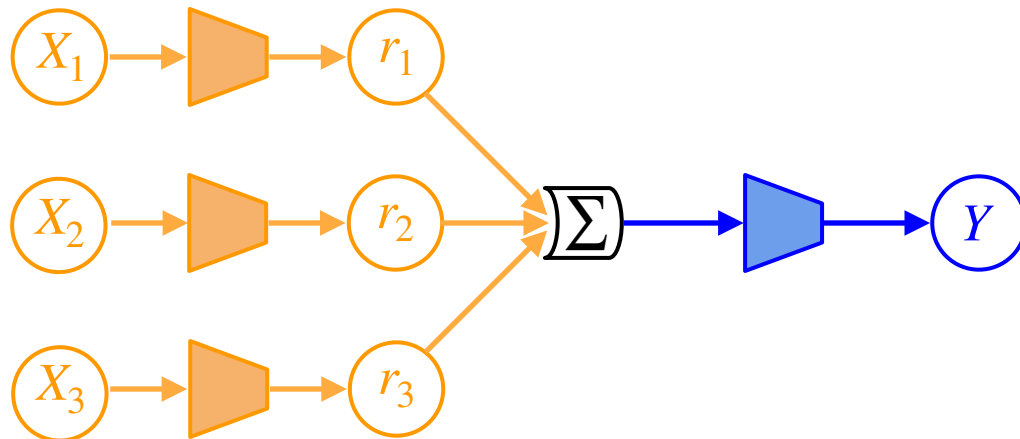
Probabilistic Permutation-Invariance



Probabilistic Permutation-Invariance



Functional Permutation-Invariance



Probabilistic Permutation-Equivariance

- Now suppose we have random sequences $\mathbf{X}_n$ and $\mathbf{Y}_n$.
 - $\mathbf{Y}_n$ is conditionally permutation-equivariant given $\mathbf{X}_n$.
 - $\mathbf{X}_n$ is marginally permutation-invariant (exchangeable).
- Also suppose that $Y_i \perp\!\!\!\perp \mathbf{Y}_n \setminus Y_i \mid \mathbf{X}_n$ for each i .

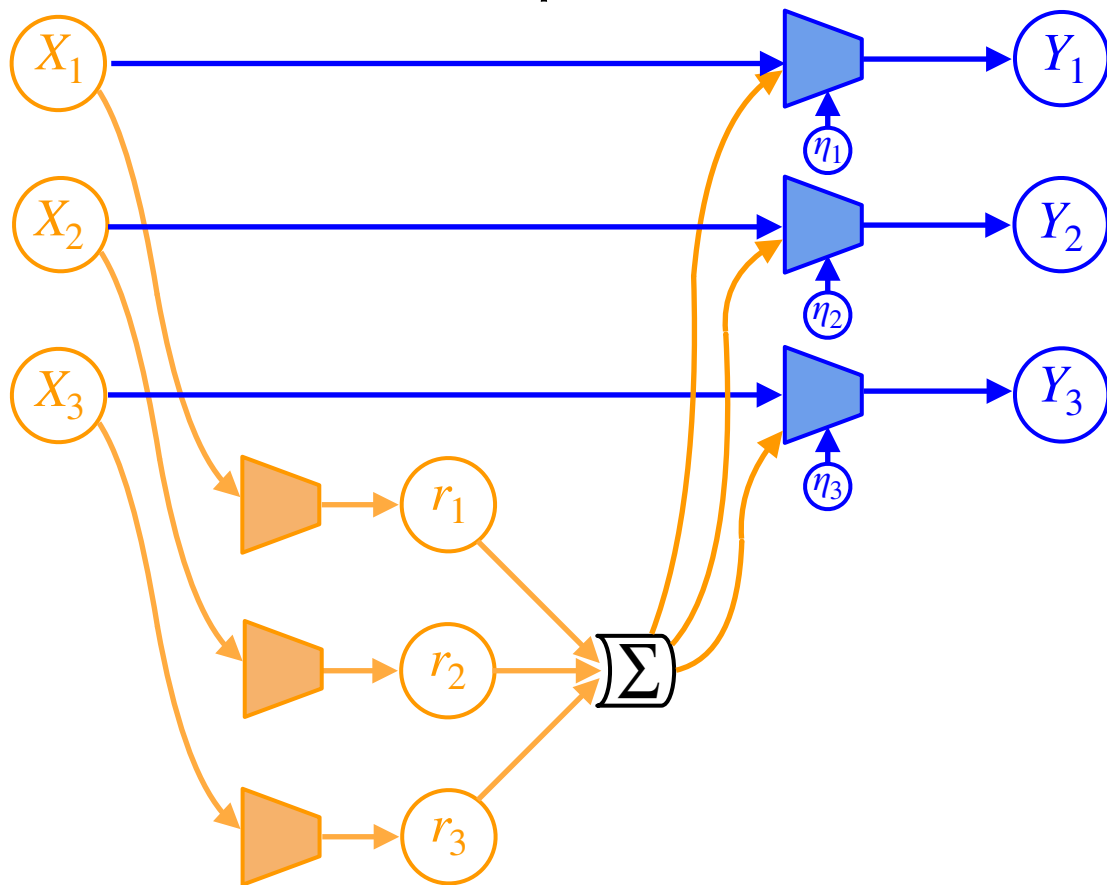
- Then:

$$(\mathbf{X}_n, (Y_1, \dots, Y_n)) =_{a.s.} (\mathbf{X}_n, (h(\eta_1, X_1, \mathbb{M}_{\mathbf{X}_n}), \dots, h(\eta_n, X_n, \mathbb{M}_{\mathbf{X}_n})))$$

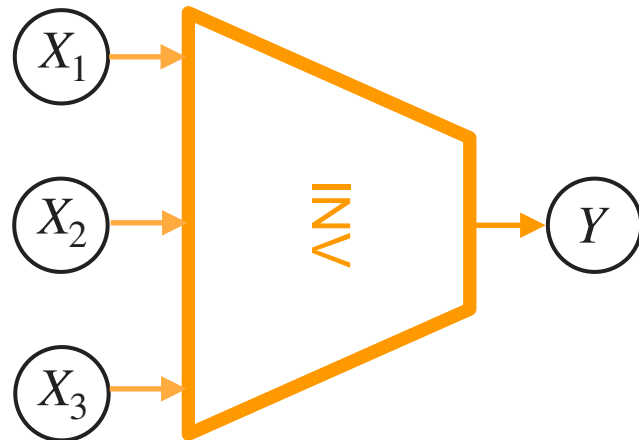
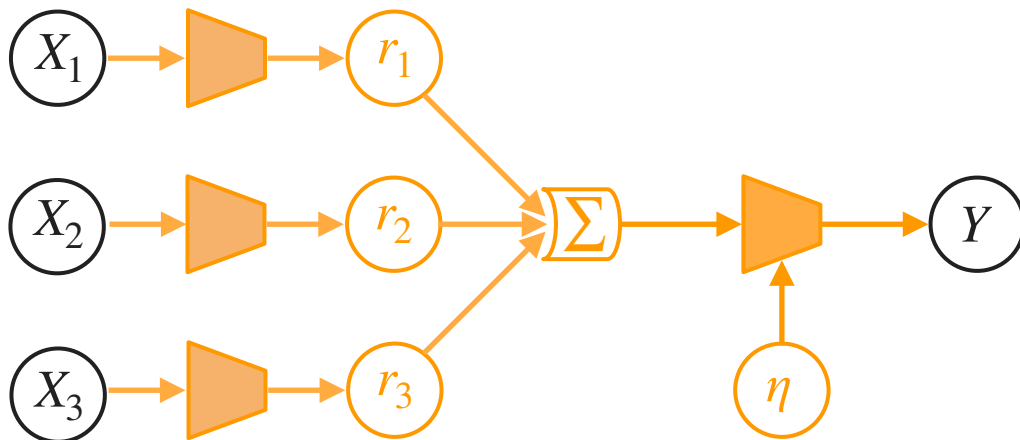
for outsourced noise (η_i) that is mutually independent and independent of $\mathbf{X}_n$.



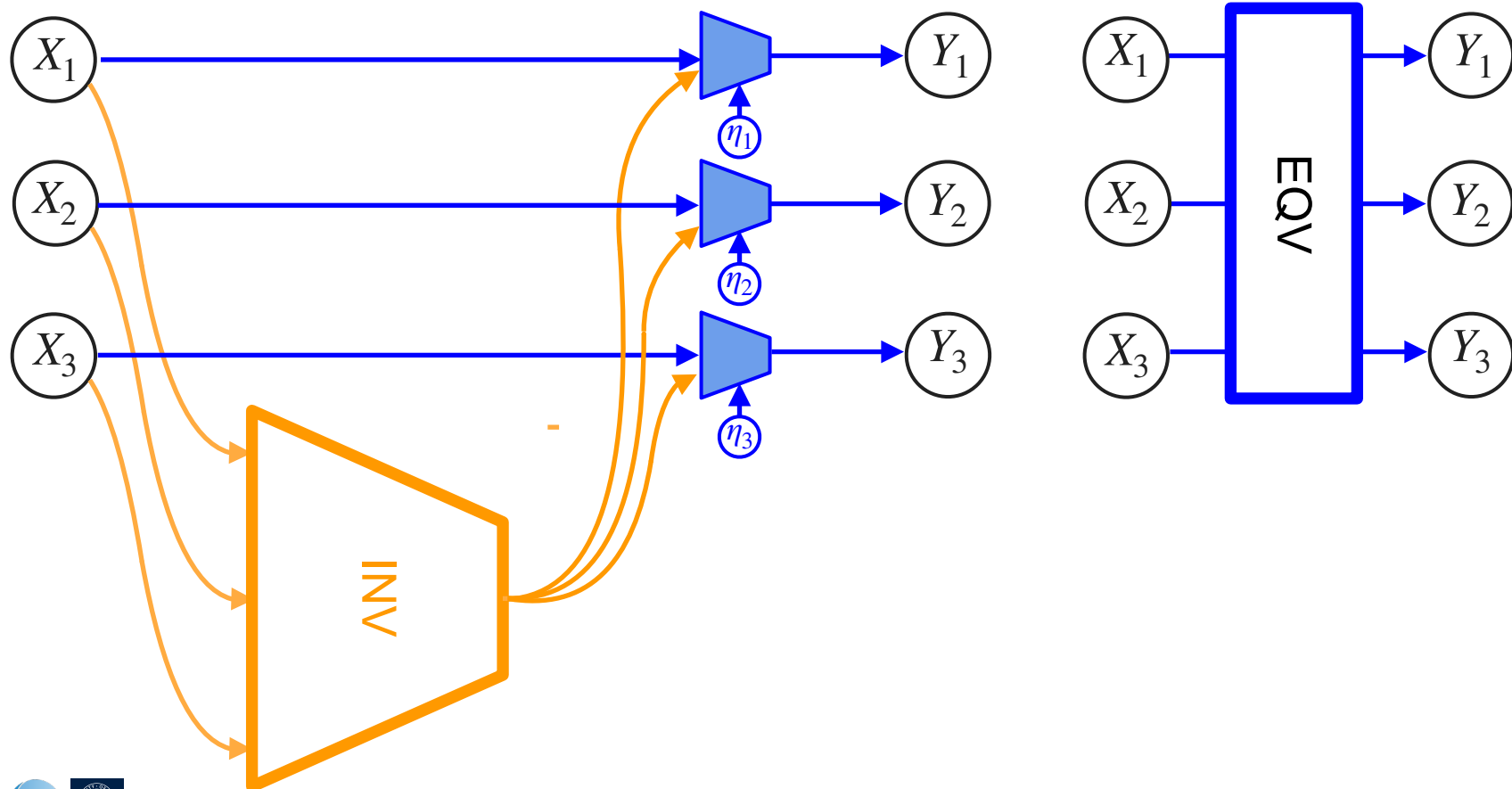
Probabilistic Permutation-Equivariance



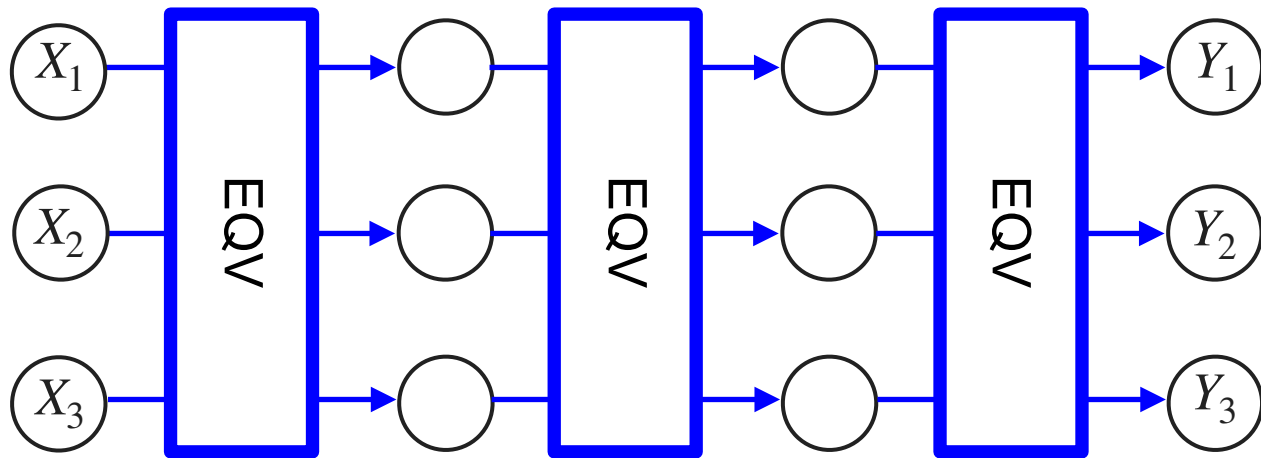
Composing Invariant and Equivariant Modules



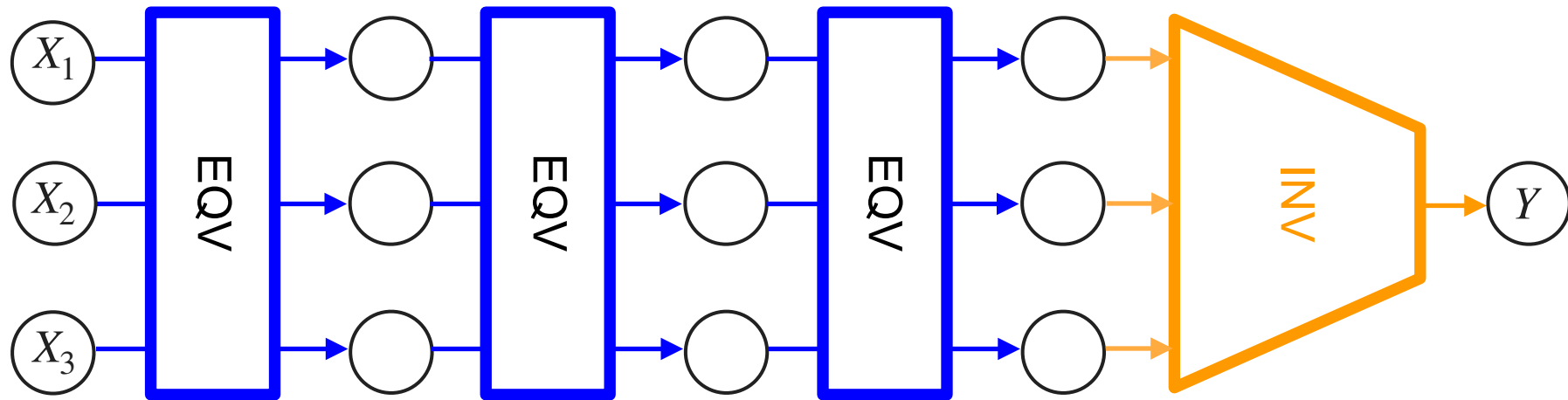
Composing Invariant and Equivariant Modules



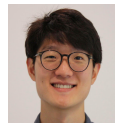
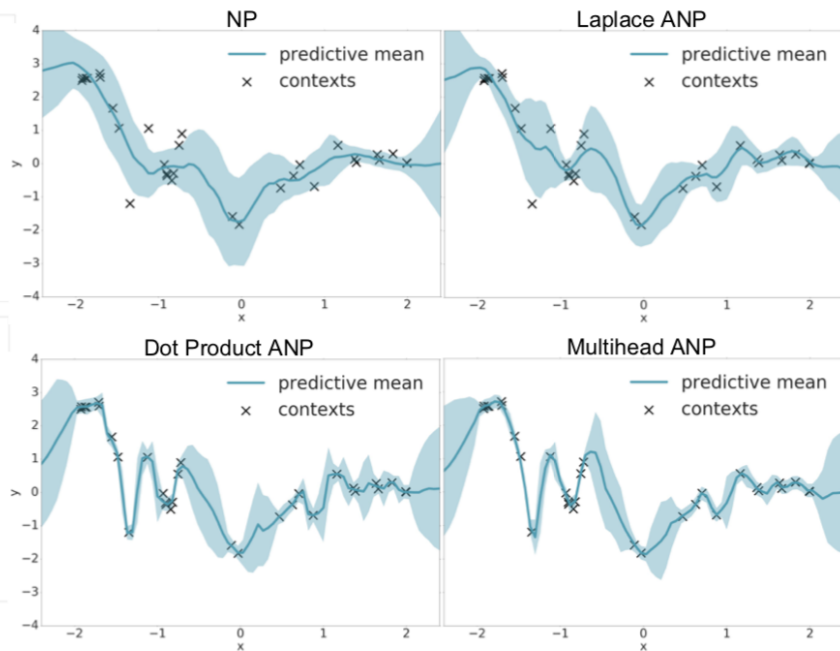
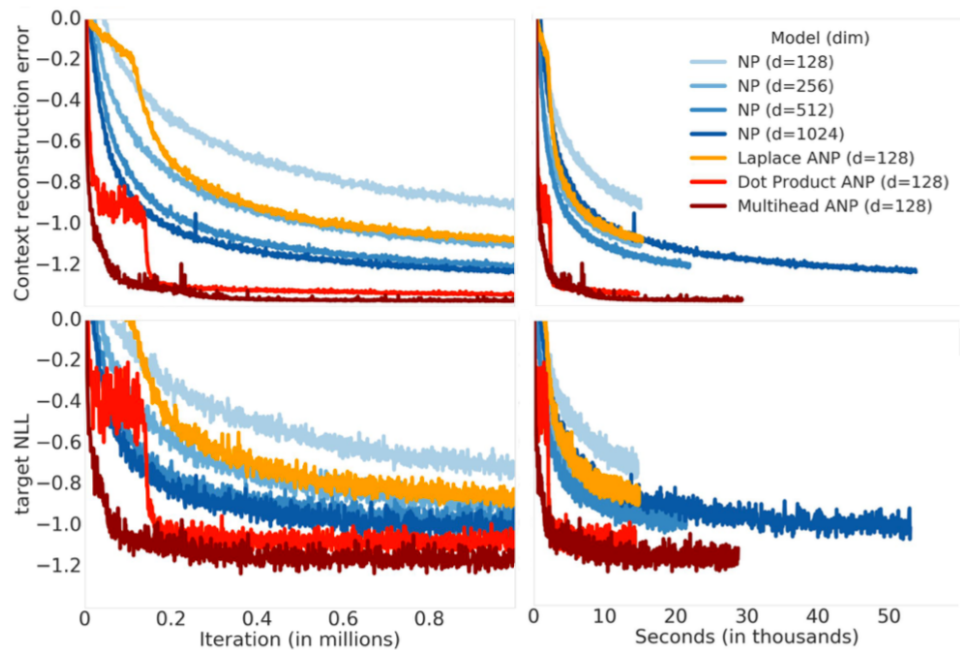
Composing Invariant and Equivariant Modules



Composing Invariant and Equivariant Modules



Attentive Neural Processes



Maximal Invariant and Maximal Equivariant

- Let G be a compact group.
- A maximal invariant is a statistic $M : \mathcal{X} \mapsto \mathcal{S}$ such that

$$M(g \cdot x) = M(x) \forall g \in G, x \in \mathcal{X}$$

$$M(x_1) = M(x_2) \Rightarrow \exists g \in G : x_1 = g \cdot x_2$$

- A maximal equivariant $\tau : \mathcal{X} \mapsto G$ satisfies

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



Probabilistic and Functional Symmetries

- Let G be a compact group and X be marginally G -invariant.
- Let M be a maximal invariant, then
 Y is conditionally G -invariant given $X \Leftrightarrow (X, Y) =_{a.s.} (X, h(\eta, M(X)))$
for outsourced noise η independent of X and a function h .
- If a maximal equivariant τ exists and $G_X \subset G_Y$ a.s., then
 Y is conditionally G -equivariant given $X \Leftrightarrow (X, Y) =_{a.s.} (X, h(\eta, X))$
for outsourced noise η independent of X and a function h that is G -equivariant in its second argument.



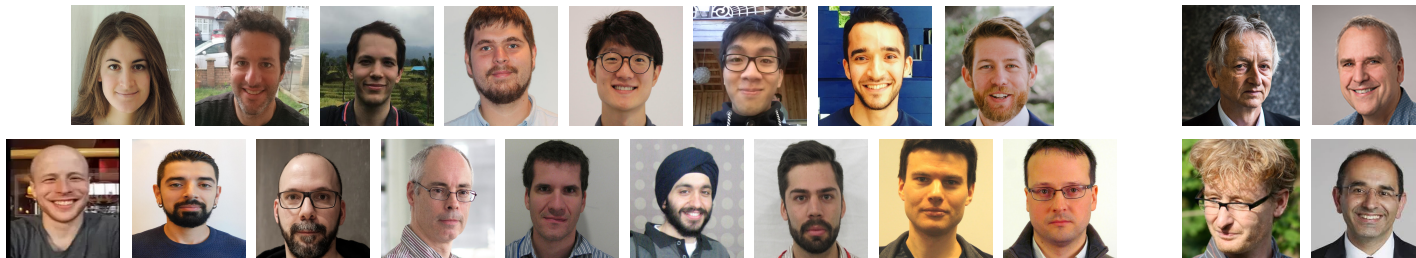
Concluding Remarks

- Neural processes allow us to learn domain-specific priors from data.
 - Focussing on predictive (conditional) distributions with latent processes marginalised out works very well [Garnelo et al 2018b, Le et al 2018].
 - Learnt conditional distributions are not guaranteed to be consistent.
- Tools from probabilities symmetries, sufficiency and adequacy allowed us to answer questions about neural architectures under symmetry.
 - Framework extends to graph and array structured data with node exchangeability.
 - How to relax assumptions of conditional independence of outputs?



Thank You!

- Collaborators, colleagues, mentors



- Openings @ Oxford: <http://www.stats.ox.ac.uk/vacancies/>
 - Director of Statistical Consultancy
 - Florence Nightingale Bicentennial Fellowship (5 year “super-postdocs”)

