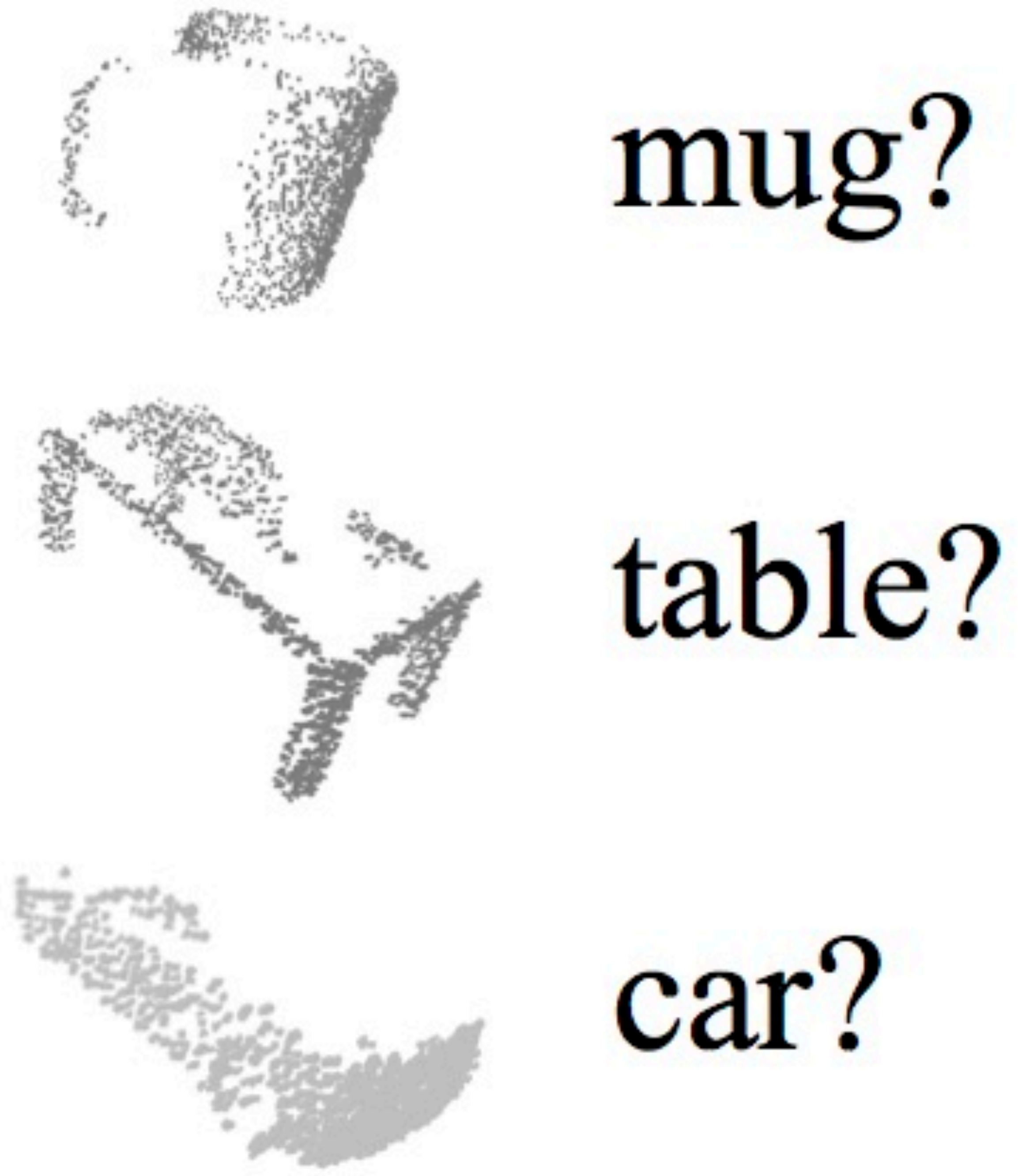
Set Norm and Equivariant Skip Connections: Putting the Deep in Deep Sets

Lily H. Zhang*, Veronica Tozzo*, John M. Higgins, Rajesh Ranganath

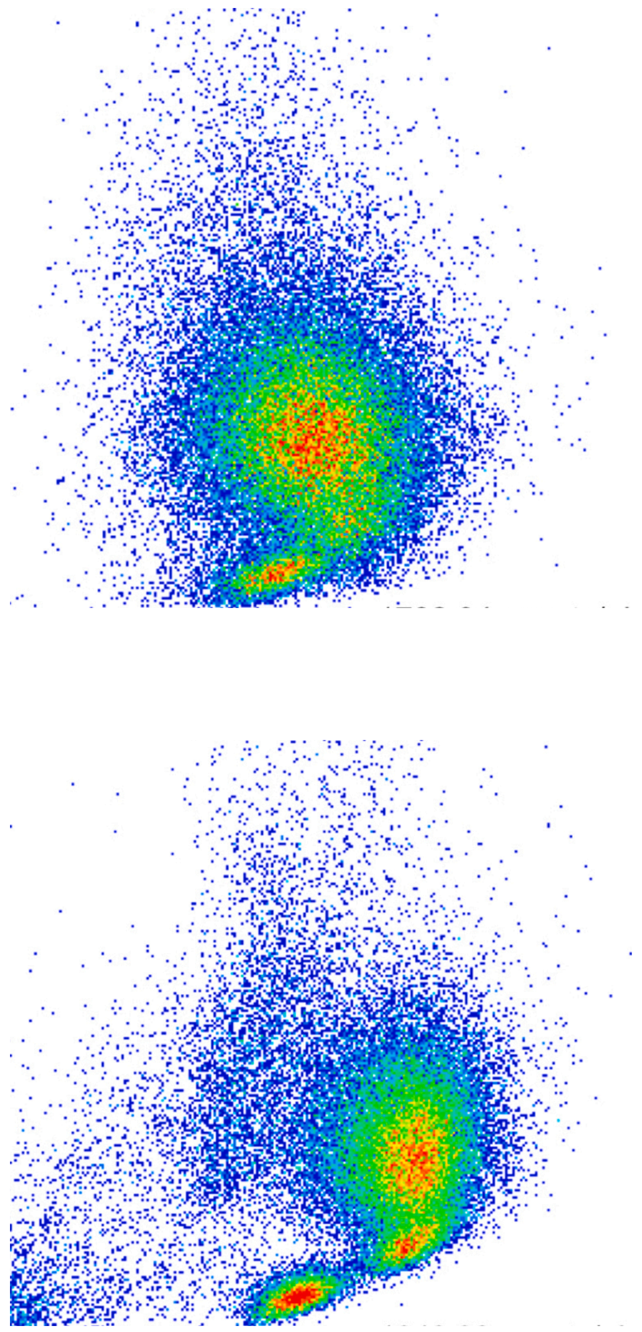
NYU

MGH 1811 MASSACHUSETTS GENERAL HOSPITAL PATHOLOGY

VE RI TAS HARVARD MEDICAL SCHOOL



Point cloud classification

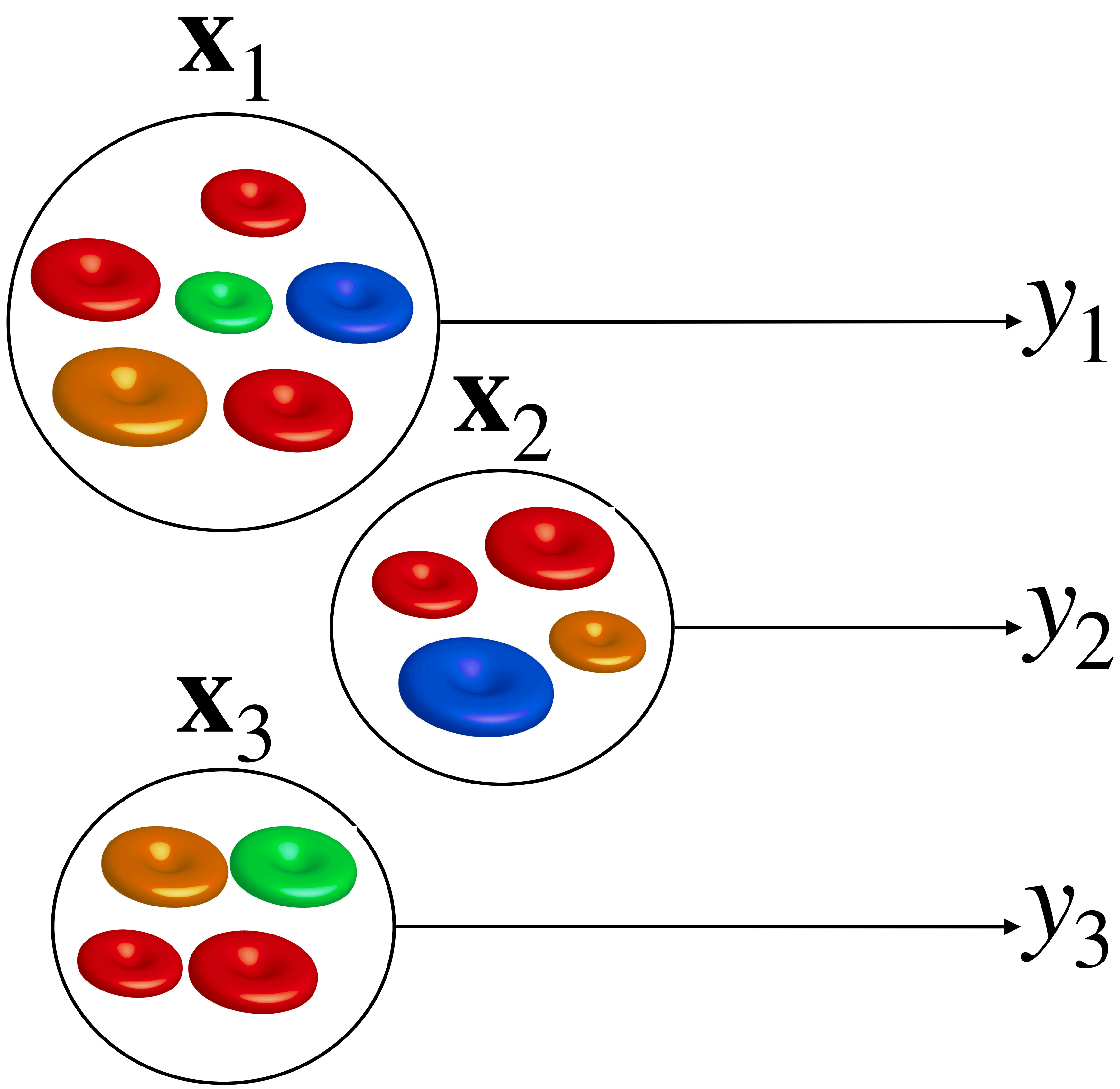


healthy?

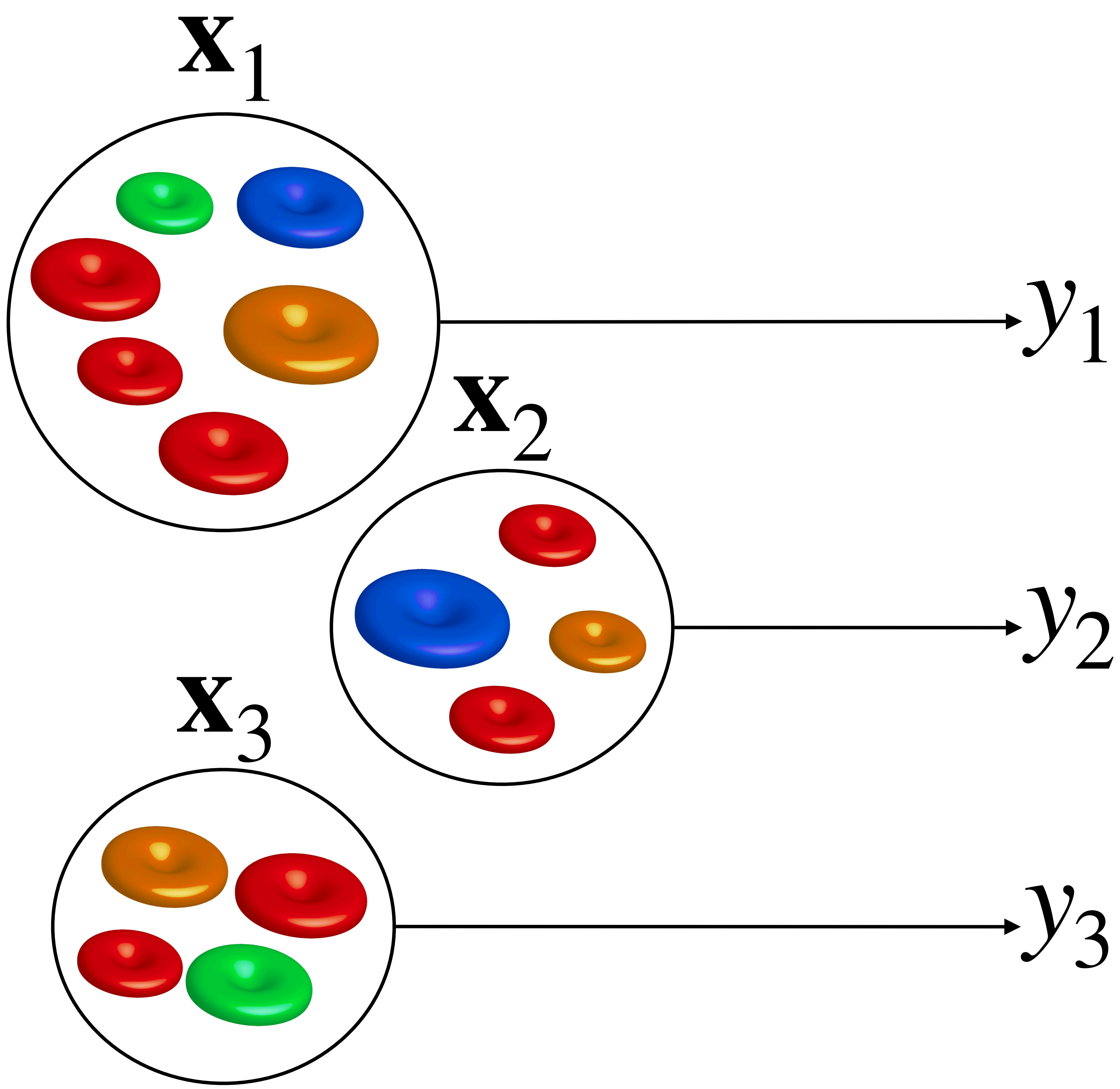
ill?

Prediction of health outcomes
from single-cell data

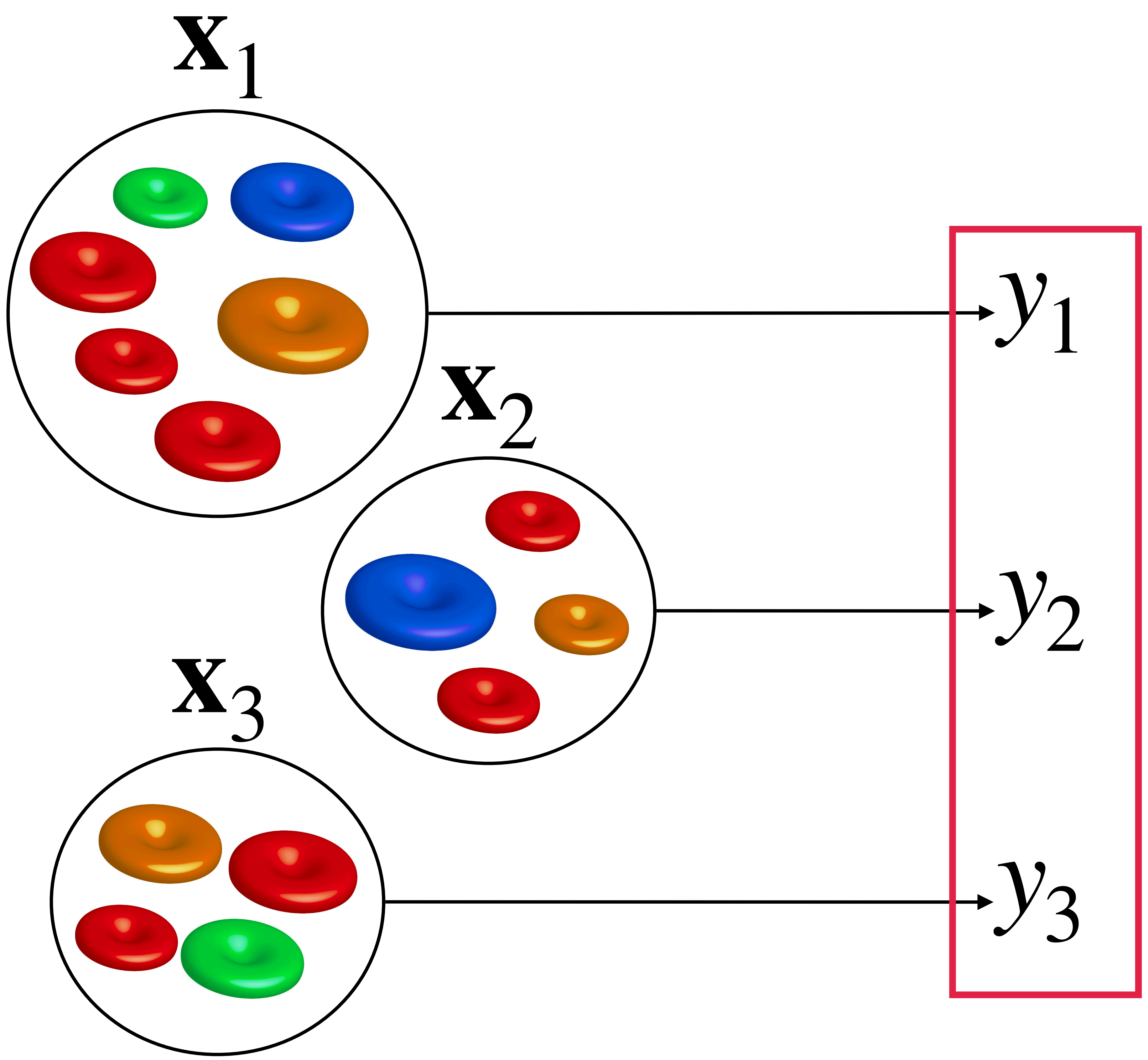
The problem



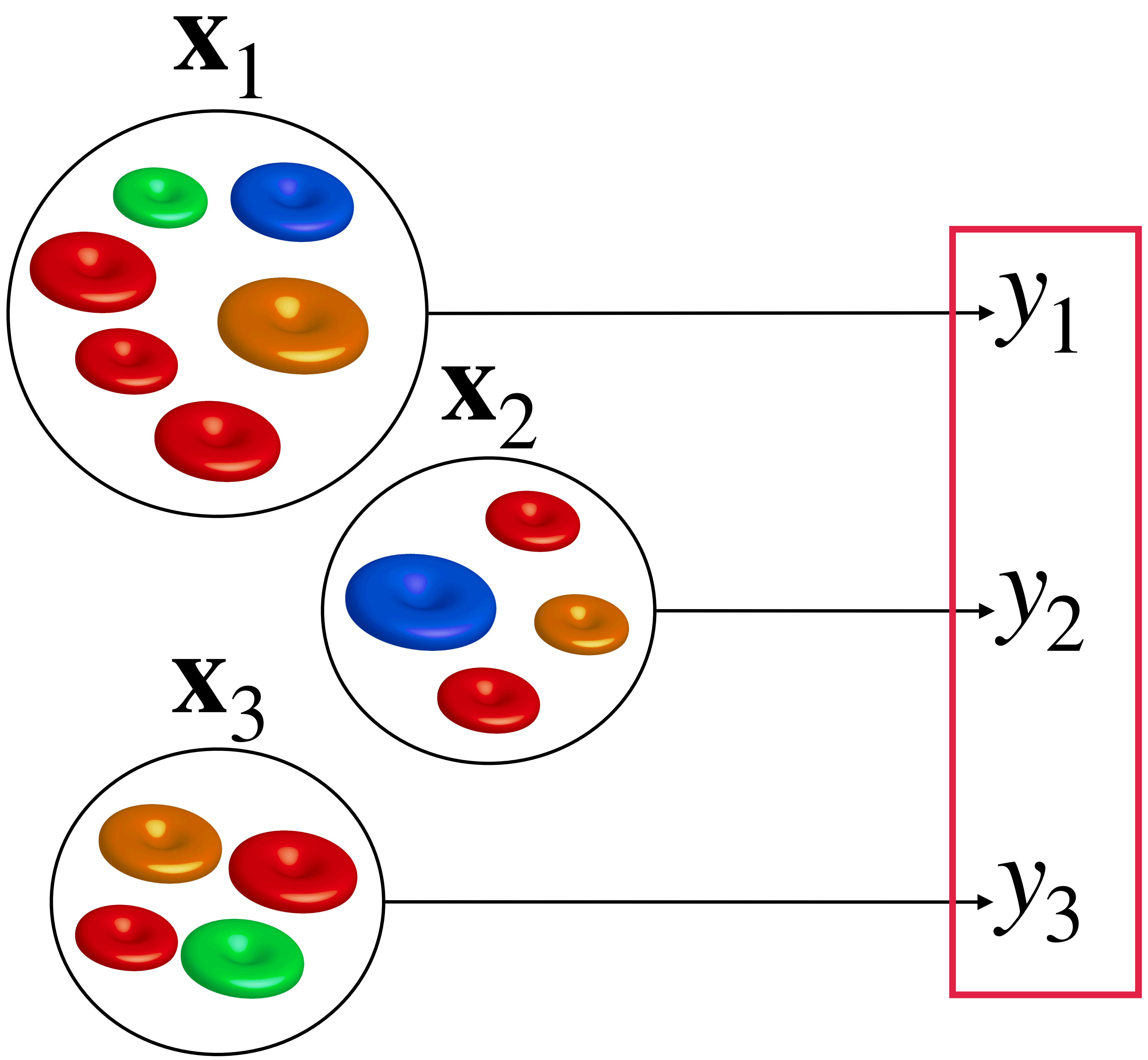
The problem



The problem



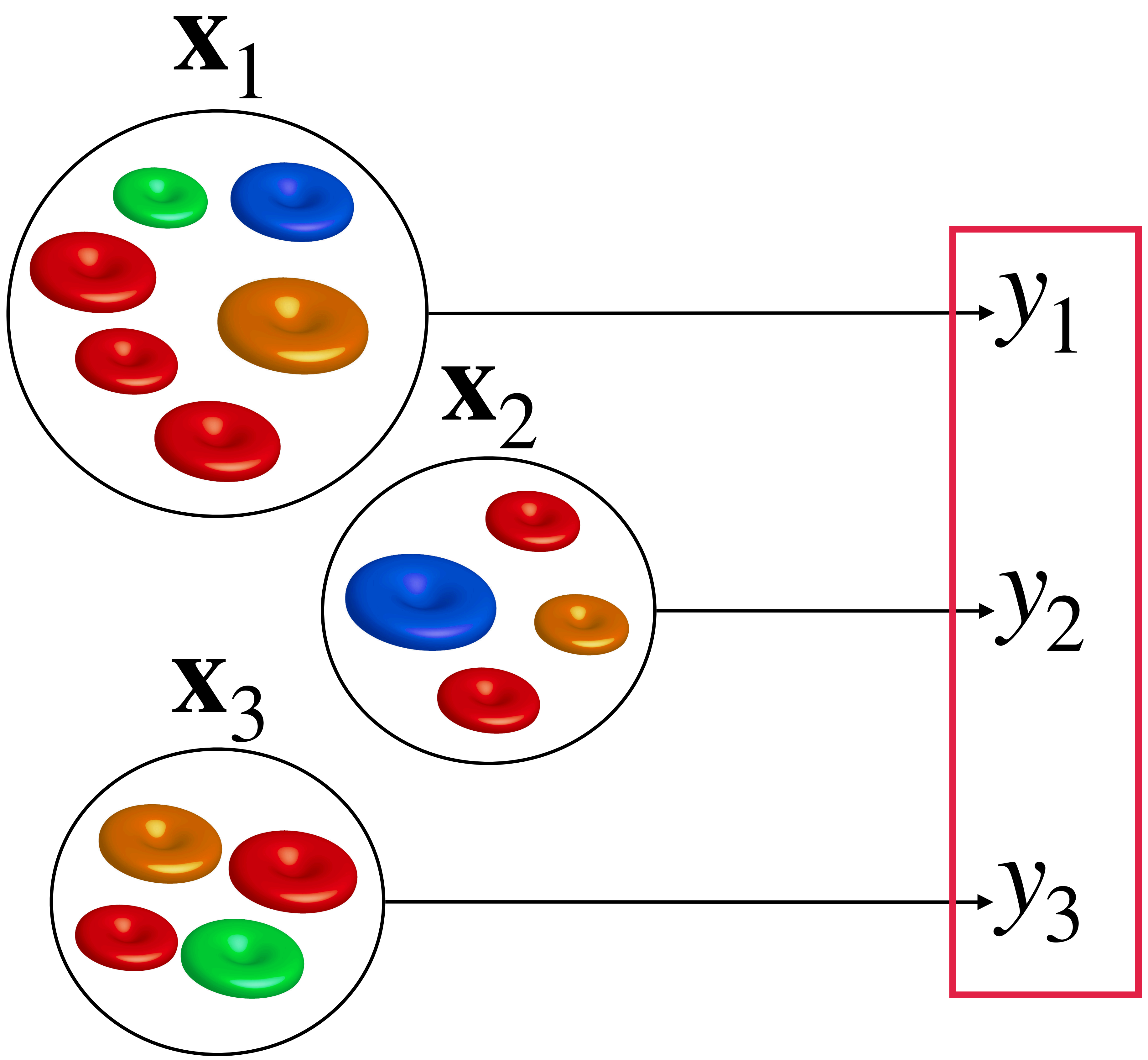
The problem



State-of-the-art methods:

- Deep Sets (Zaheer et al. 2018)
- Set Transformer (Lee et al. 2019)

The problem



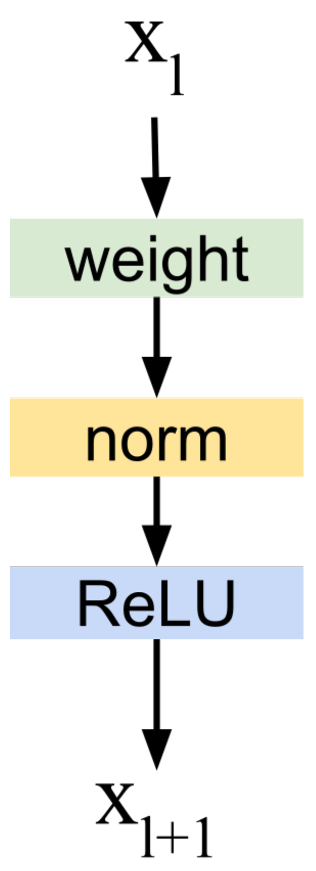
State-of-the-art methods:

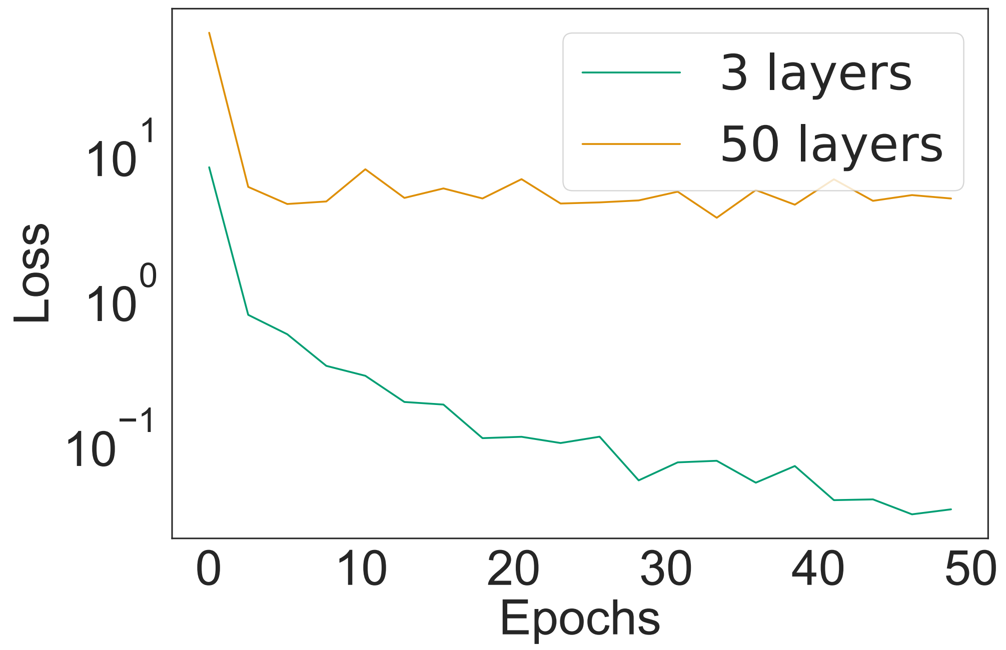
- Deep Sets (Zaheer et al. 2018)
- Set Transformer (Lee et al. 2019)

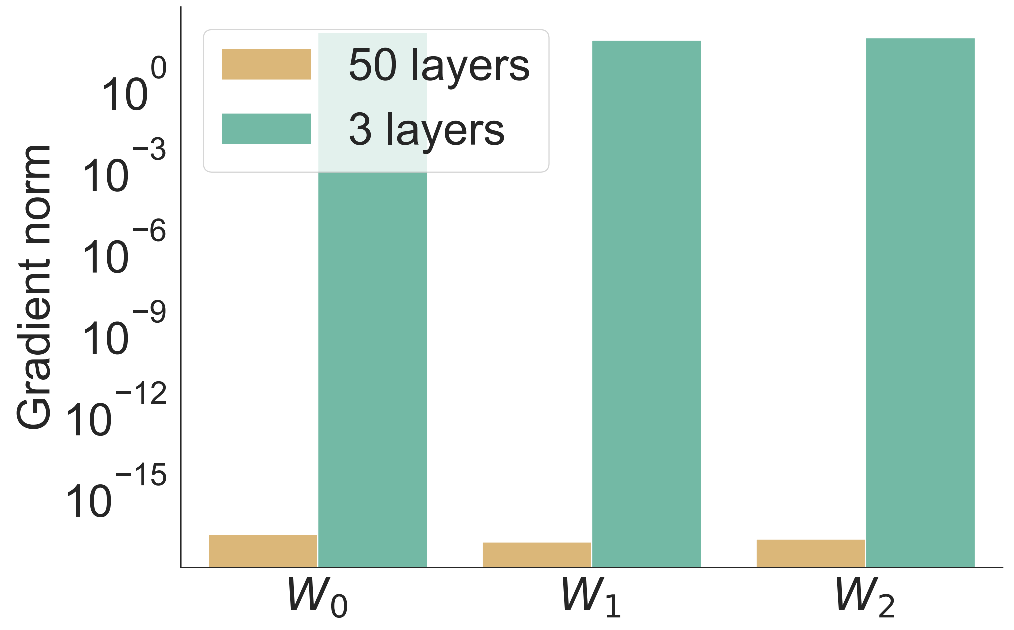
What happens when we go deep?

Deep Sets and Set Transformer suffer from vanishing/exploding gradients

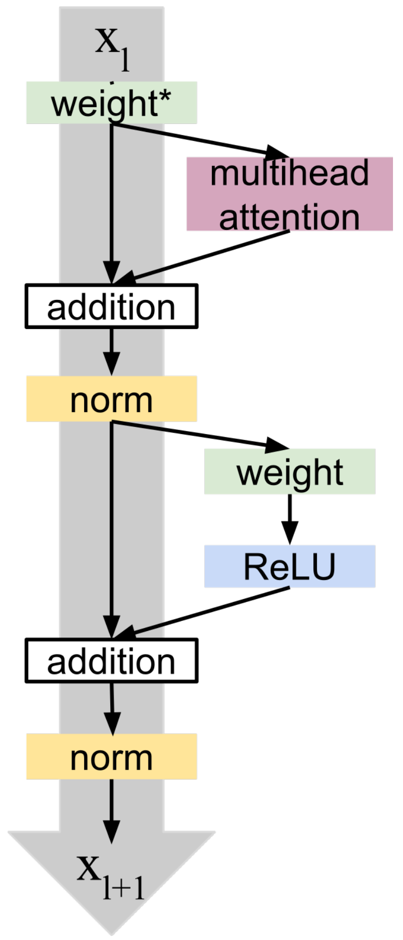
Deep Sets layer

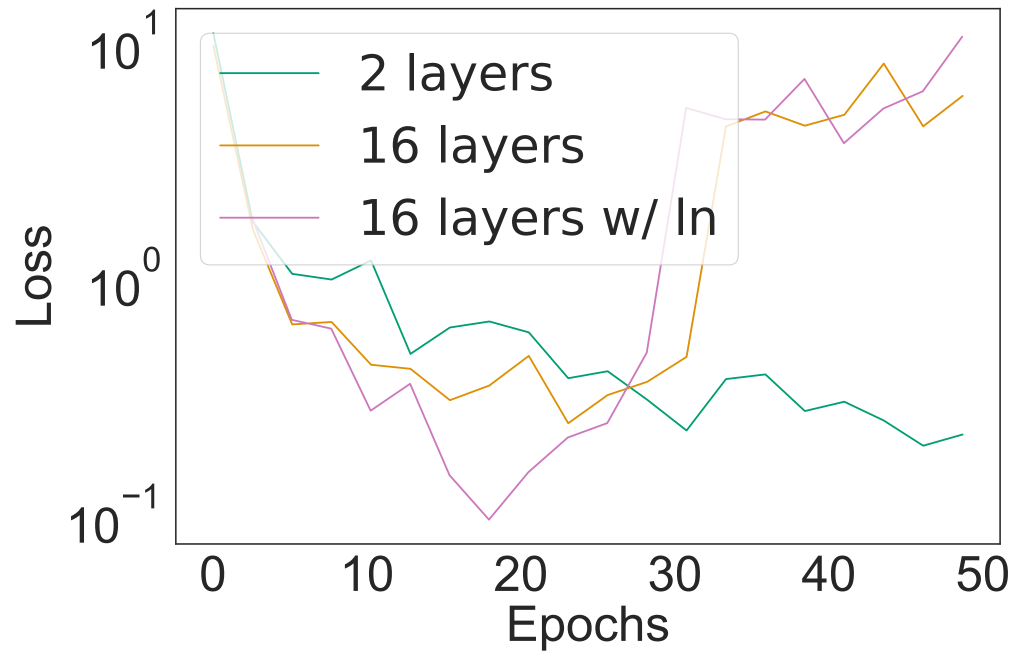


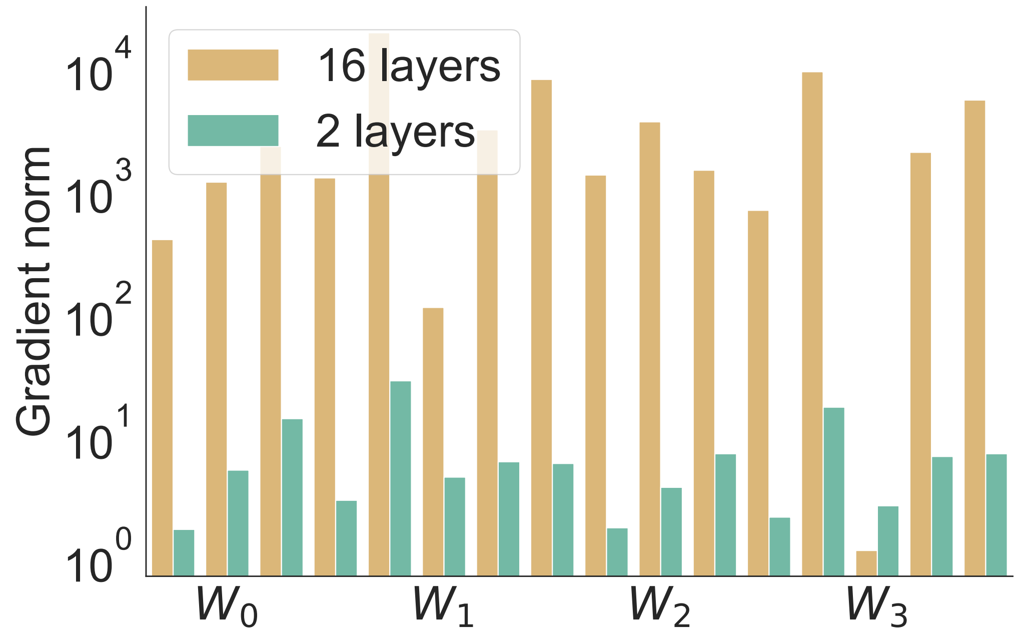




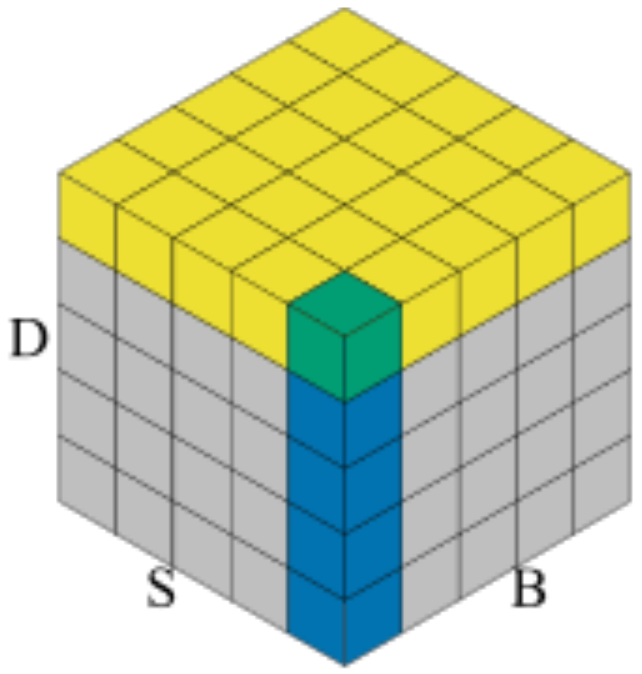
Set Transformer layer





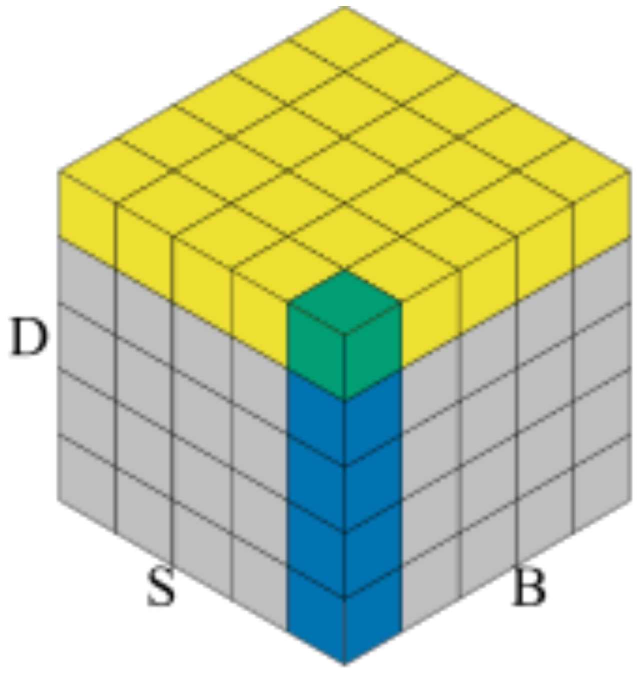


Layer Norm forces unwanted invariances



Layer norm
Per set, per sample standardization
Per feature transformation

Layer Norm forces unwanted invariances



Layer norm
Per set, per sample standardization
Per feature transformation

$$\mathbf{x}_m = \alpha \mathbf{x}_{m'} \Rightarrow \mathrm{LN}(\mathbf{x}_s W) = \mathrm{LN}(\mathbf{x}_{s'} W)$$

Deep Sets ++ and Set Transformer ++

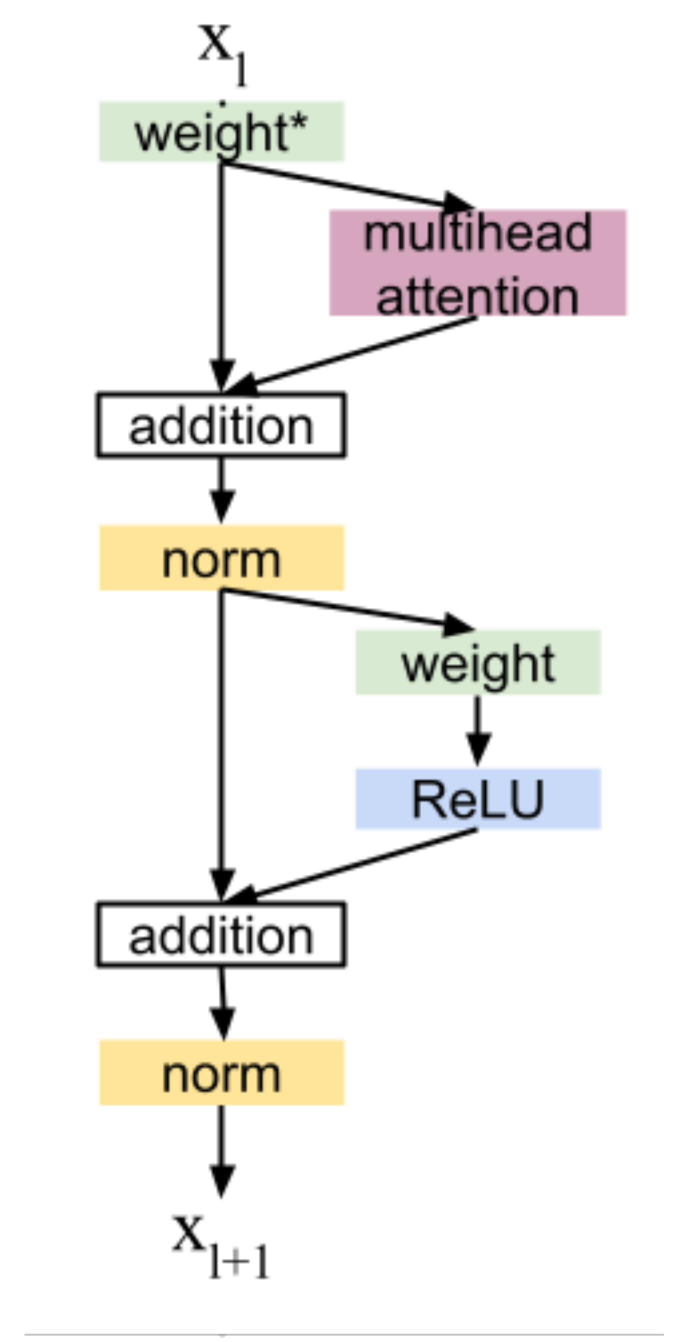
Deep Sets ++ and Set Transformer ++

- Careful design of residual connections, **Clean path residual connections**
- Normalization layer specific for sets, **Set Norm**

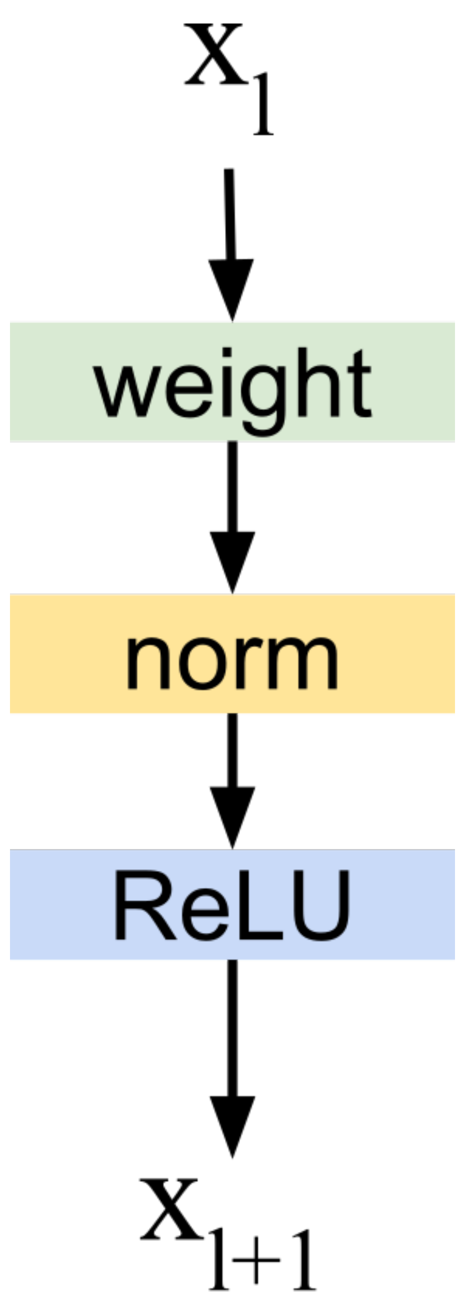
Deep Sets ++ and Set Transformer ++

- Careful design of residual connections, **Clean path residual connections**
- Normalization layer specific for sets, **Set Norm**

Set Transformer



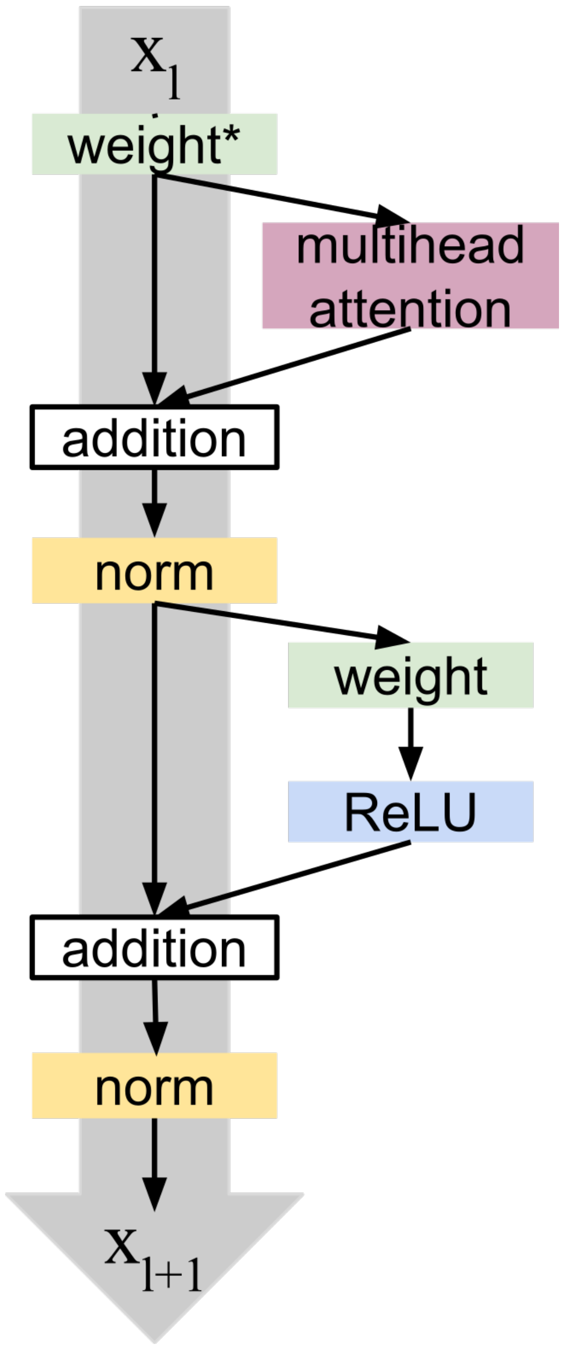
Deep Sets



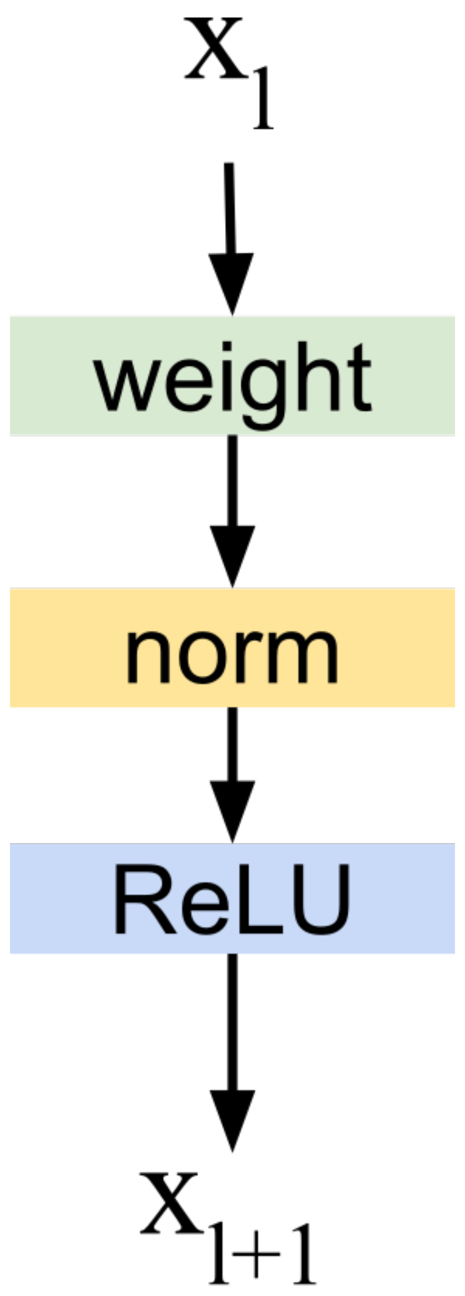
Deep Sets ++ and Set Transformer ++

- Careful design of residual connections, **Clean path residual connections**
- Normalization layer specific for sets, **Set Norm**

Set Transformer



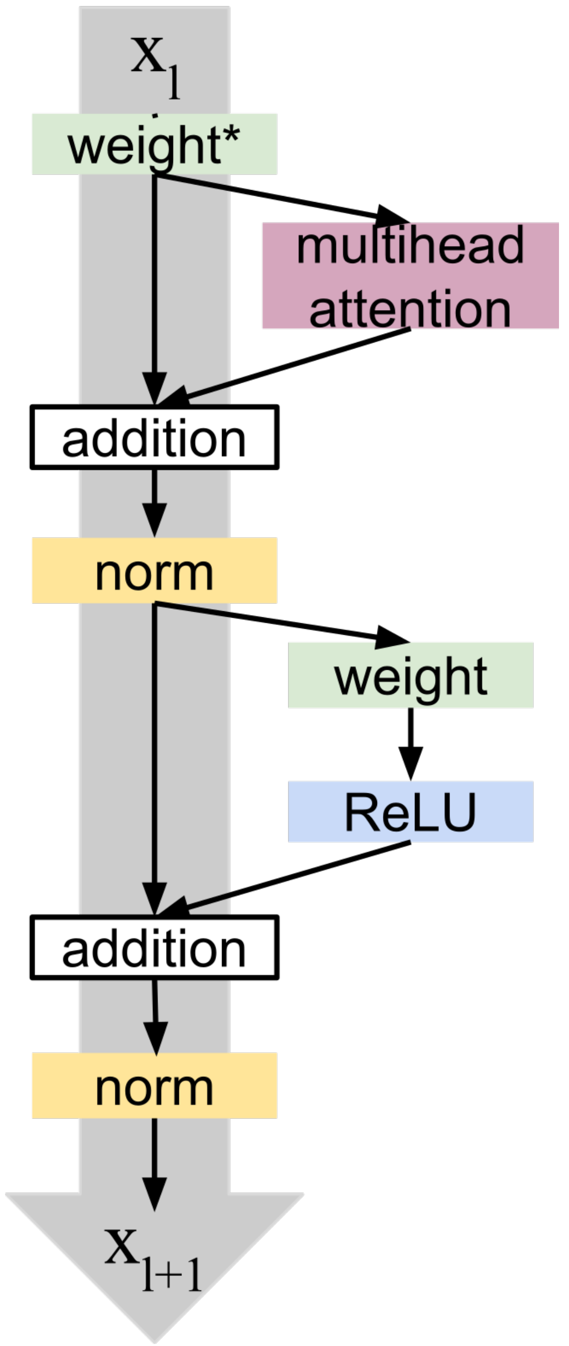
Deep Sets



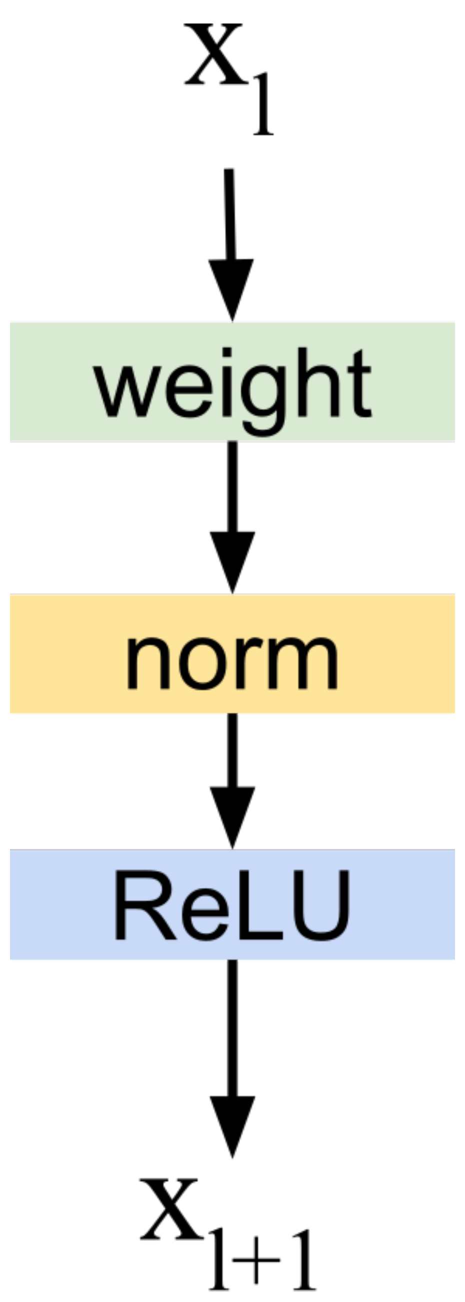
Deep Sets ++ and Set Transformer ++

- Careful design of residual connections, **Clean path residual connections**
- Normalization layer specific for sets, **Set Norm**

Set Transformer



Deep Sets

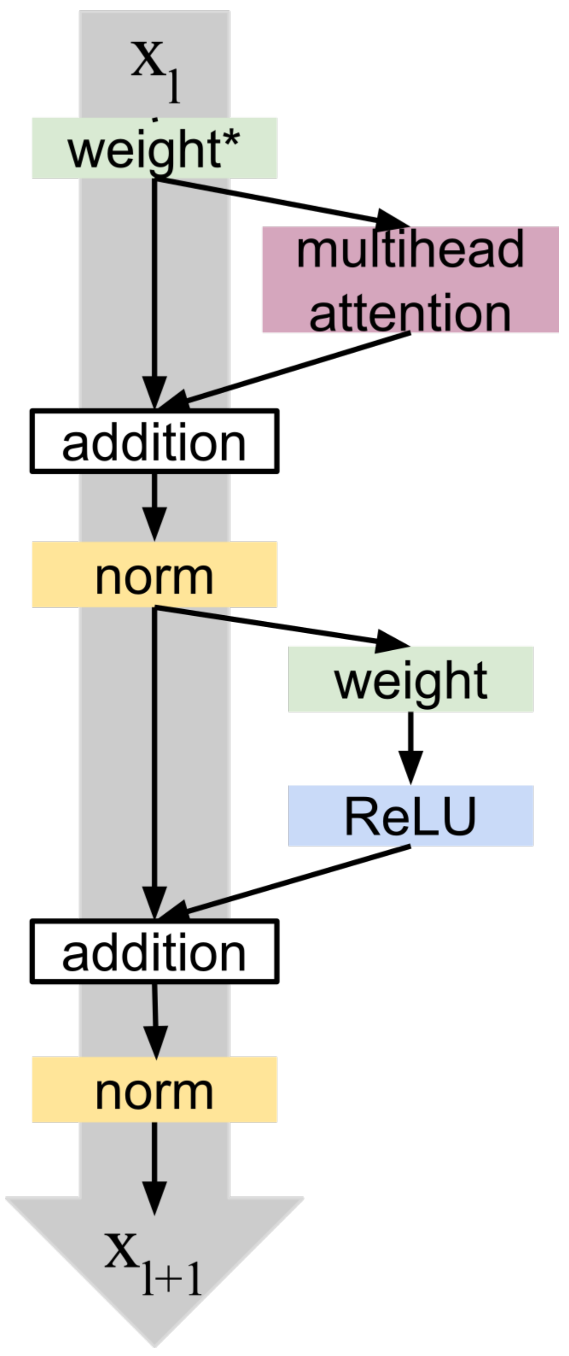


$\mathbf{x}_{l+1} = g(\mathbf{x}_l) + f(\mathbf{x}_l)$ or $\mathbf{x}_{l+1} = g(\mathbf{x}_l + f(\mathbf{x}_l))$

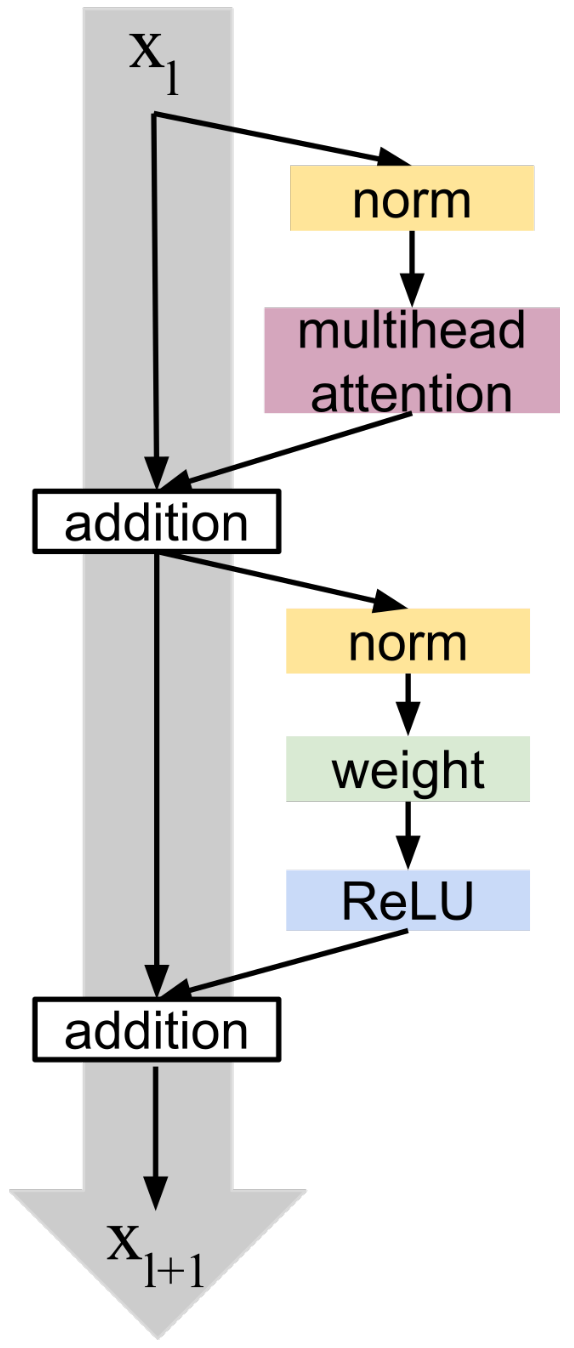
Deep Sets ++ and Set Transformer ++

- Careful design of residual connections, **Clean path residual connections**
- Normalization layer specific for sets, **Set Norm**

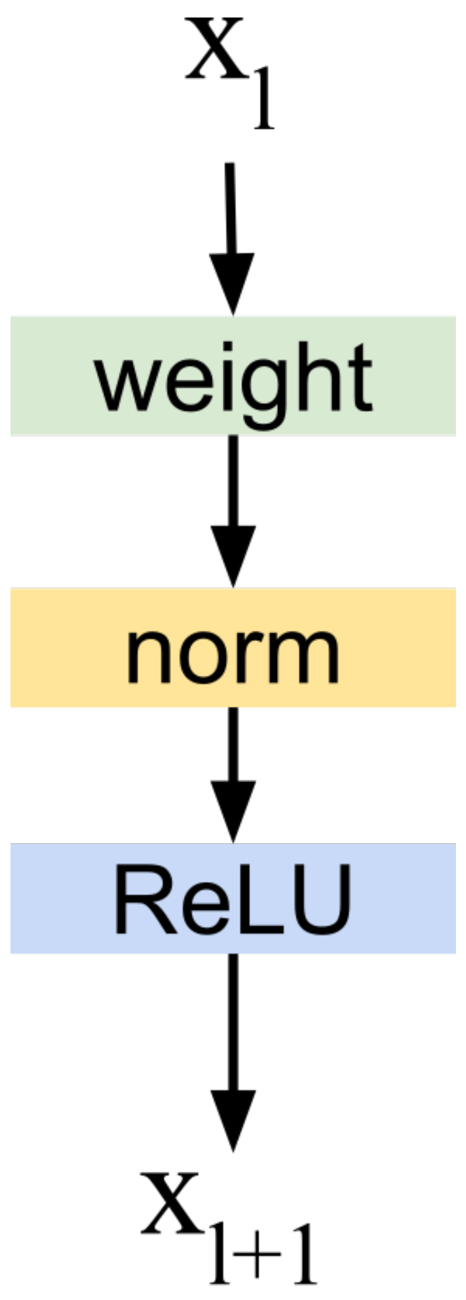
Set Transformer



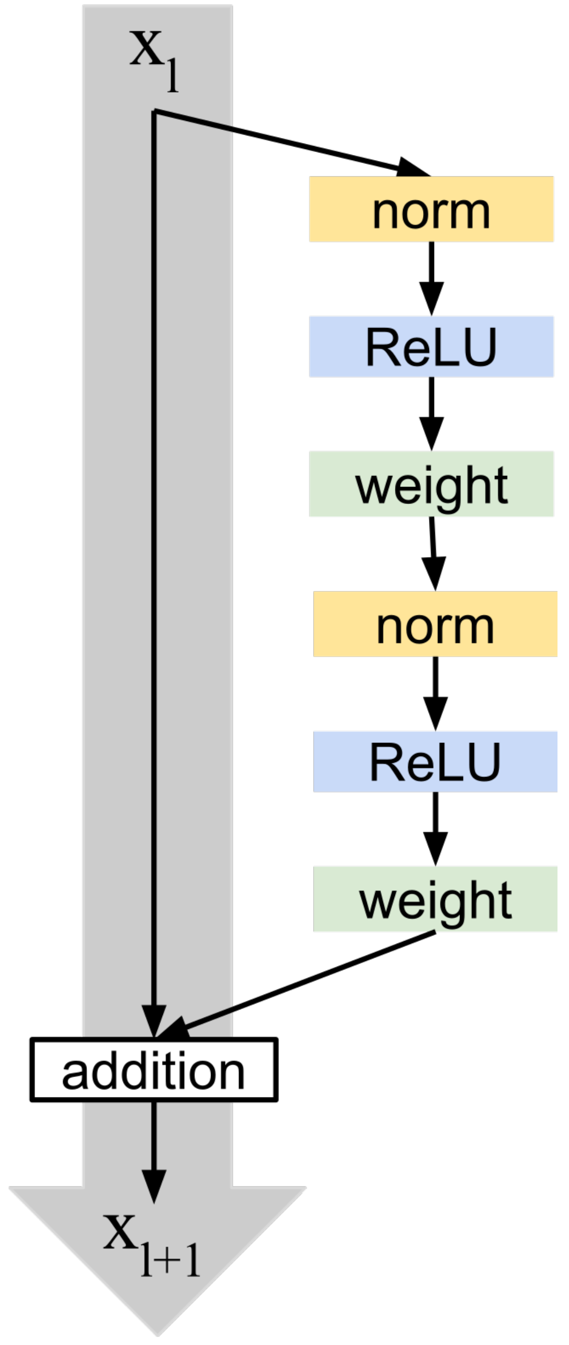
Set Transformer++



Deep Sets

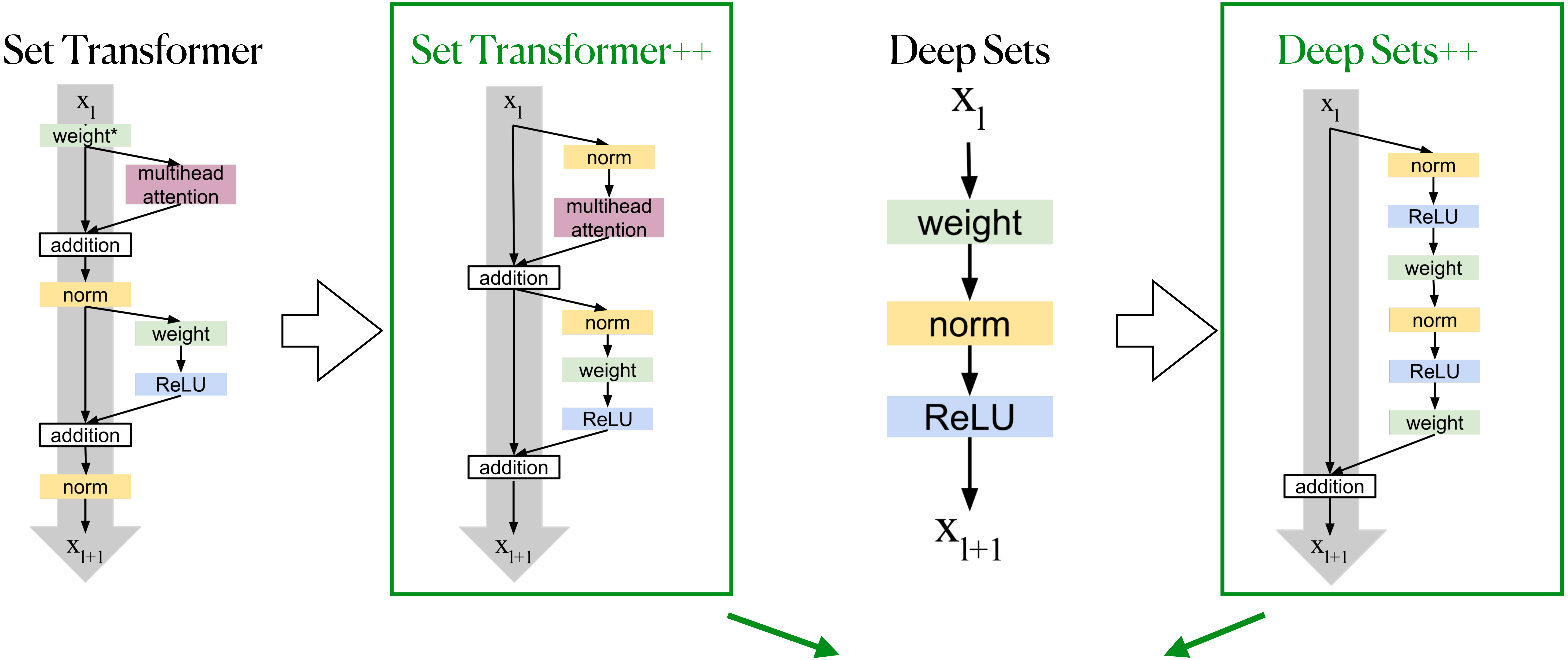


Deep Sets++



Deep Sets ++ and Set Transformer ++

- Careful design of residual connections, **Clean path residual connections**
- Normalization layer specific for sets, **Set Norm**

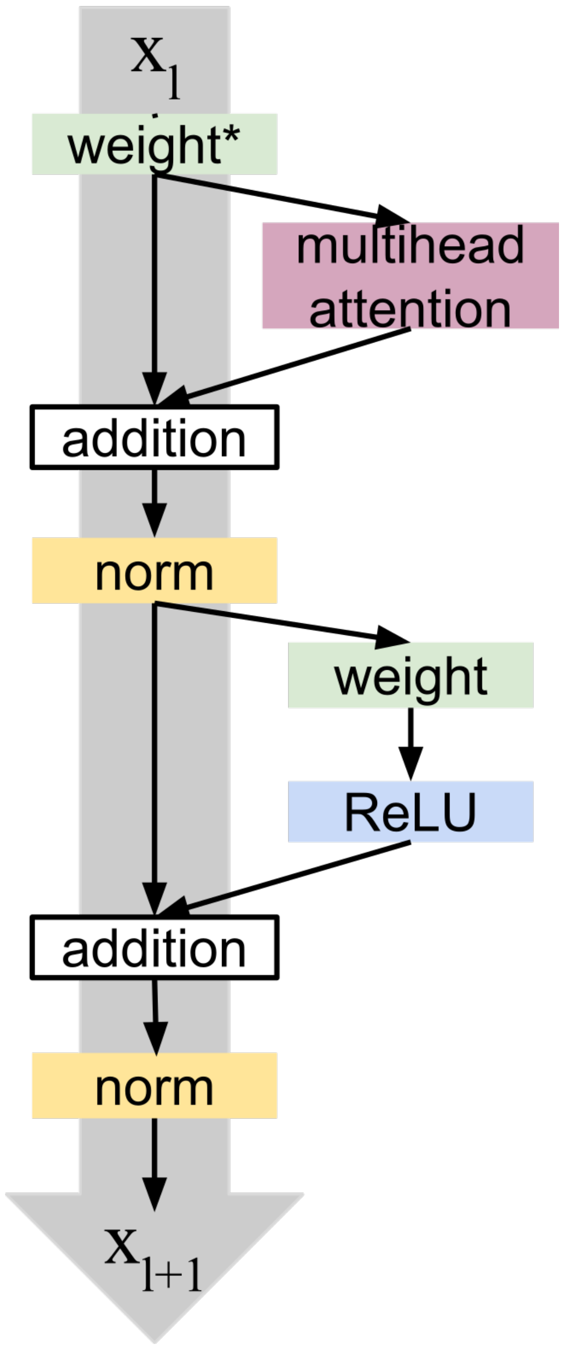


$$\mathbf{x}_{l+1} = \mathbf{x}_l + f(\mathbf{x}_l)$$

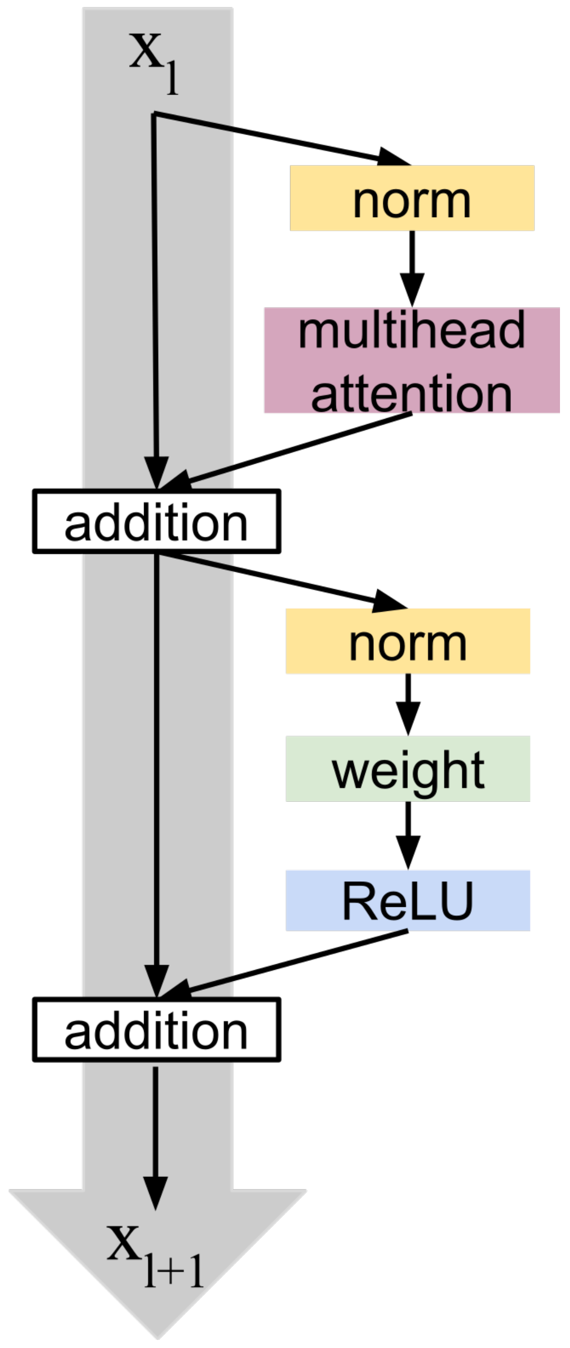
Deep Sets ++ and Set Transformer ++

- Careful design of residual connections, **Clean path residual connections**
- Normalization layer specific for sets, **Set Norm**

Set Transformer

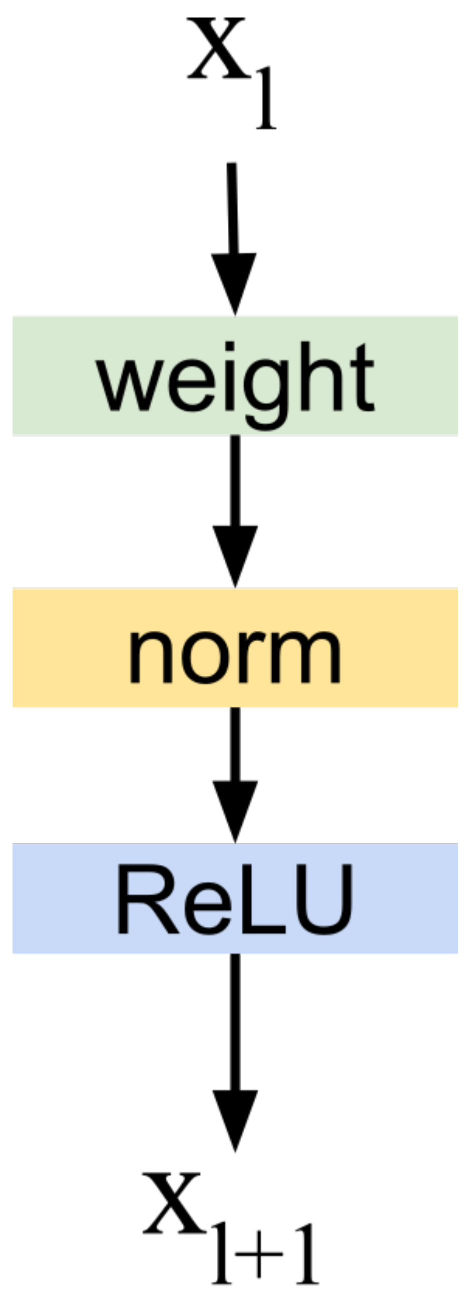


Set Transformer++

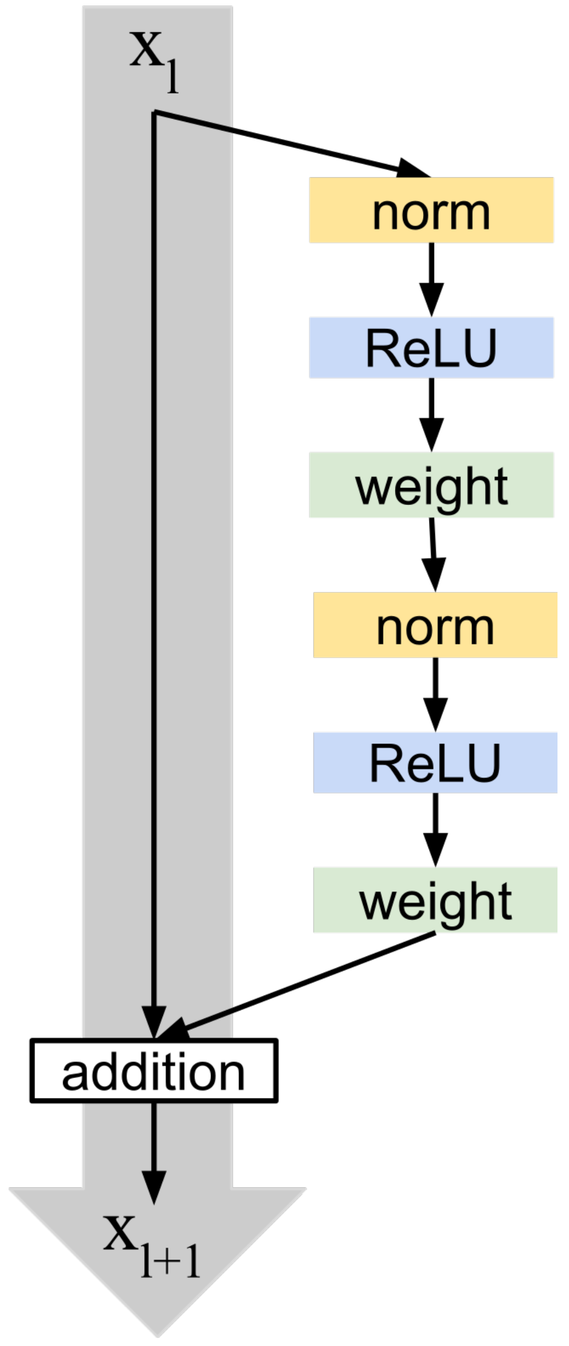


Pre-LN

Deep Sets

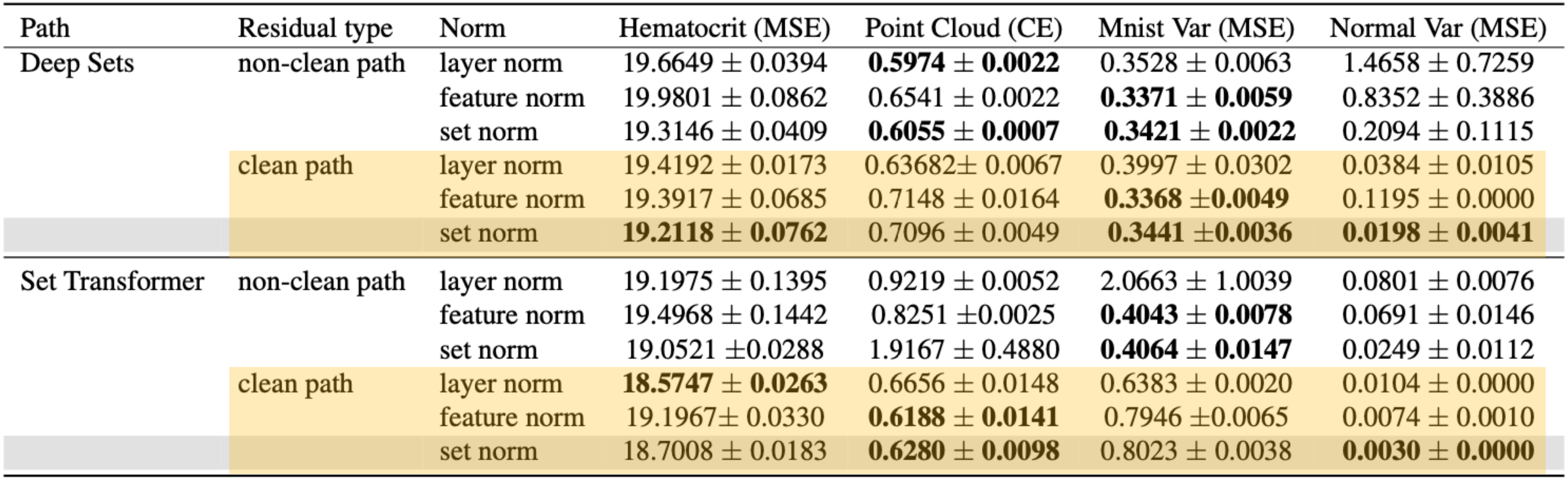


Deep Sets++



Variation of ResNet

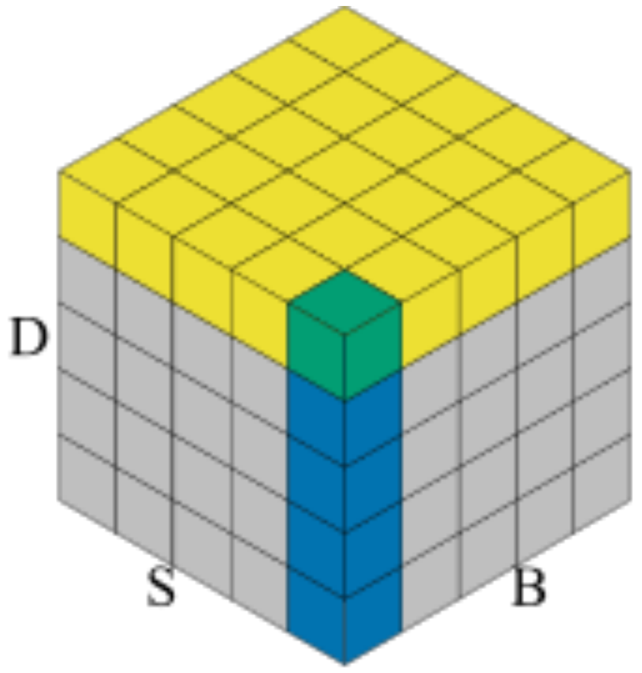
Clean path residual connections have better performances than non-clean path



Path	Residual type	Norm	Hematocrit (MSE)	Point Cloud (CE)	Mnist Var (MSE)	Normal Var (MSE)
Deep Sets	non-clean path	layer norm	19.6649 ± 0.0394	**0.5974** ± **0.0022**	0.3528 ± 0.0063	1.4658 ± 0.7259
		feature norm	19.9801 ± 0.0862	0.6541 ± 0.0022	**0.3371** ± **0.0059**	0.8352 ± 0.3886
		set norm	19.3146 ± 0.0409	**0.6055** ± **0.0007**	**0.3421** ± **0.0022**	0.2094 ± 0.1115
	clean path	layer norm	19.4192 ± 0.0173	0.63682± 0.0067	0.3997 ± 0.0302	0.0384 ± 0.0105
		feature norm	19.3917 ± 0.0685	0.7148 ± 0.0164	**0.3368** ±**0.0049**	0.1195 ± 0.0000
		set norm	**19.2118** ± **0.0762**	0.7096 ± 0.0049	**0.3441** ±**0.0036**	**0.0198** ± **0.0041**
Set Transformer	non-clean path	layer norm	19.1975 ± 0.1395	0.9219 ± 0.0052	2.0663 ± 1.0039	0.0801 ± 0.0076
		feature norm	19.4968 ± 0.1442	0.8251 ±0.0025	**0.4043** ± **0.0078**	0.0691 ± 0.0146
		set norm	19.0521 ±0.0288	1.9167 ± 0.4880	**0.4064** ± **0.0147**	0.0249 ± 0.0112
	clean path	layer norm	**18.5747** ± **0.0263**	0.6656 ± 0.0148	0.6383 ± 0.0020	0.0104 ± 0.0000
		feature norm	19.1967± 0.0330	**0.6188** ± **0.0141**	0.7946 ±0.0065	0.0074 ± 0.0010
		set norm	18.7008 ± 0.0183	**0.6280** ± **0.0098**	0.8023 ± 0.0038	**0.0030** ± **0.0000**

Deep Sets ++ and Set Transformer ++

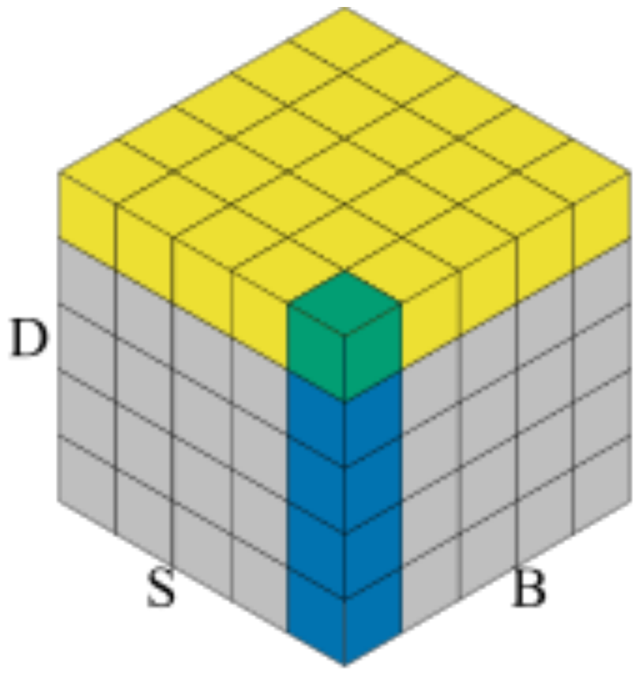
- Careful design of residual connections, **Clean path residual connections**
- Normalization layer specific for sets, **Set Norm**



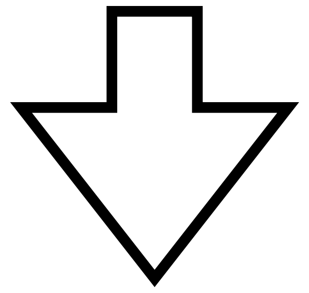
Layer norm
Per set, per sample standardization
Per feature transformation

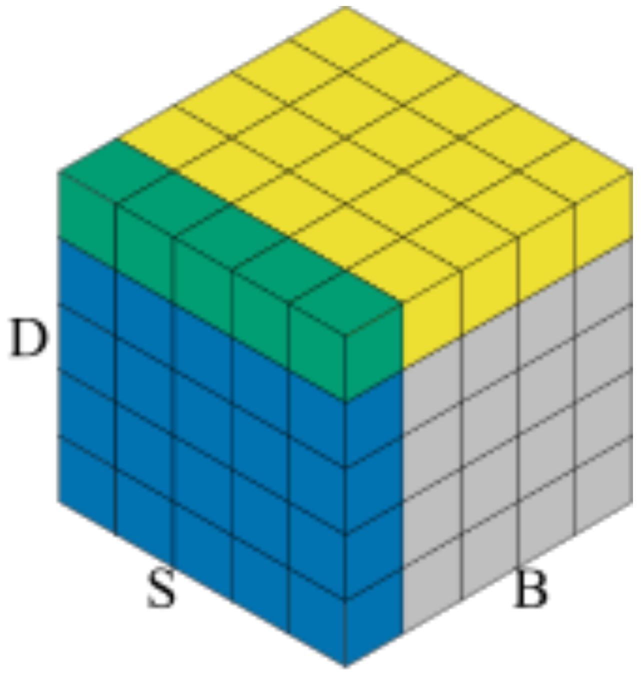
Deep Sets ++ and Set Transformer ++

- Careful design of residual connections, **Clean path residual connections**
- Normalization layer specific for sets, **Set Norm**



Layer norm
Per set, per sample standardization
Per feature transformation

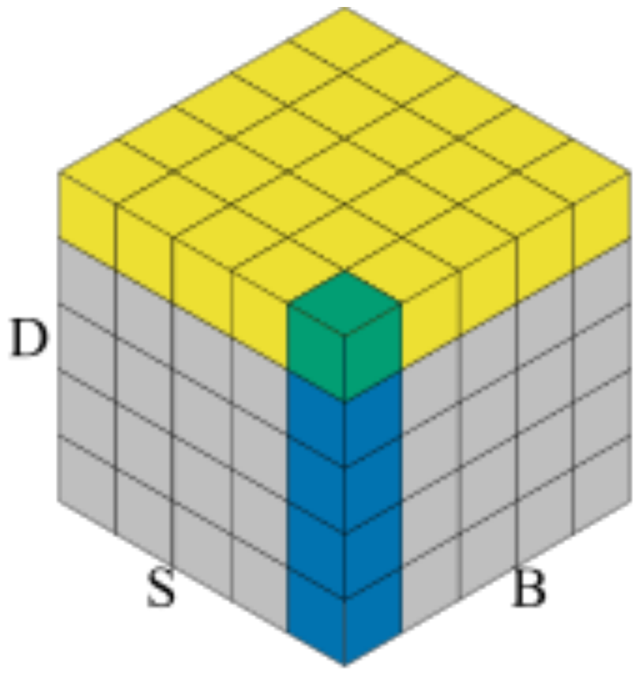




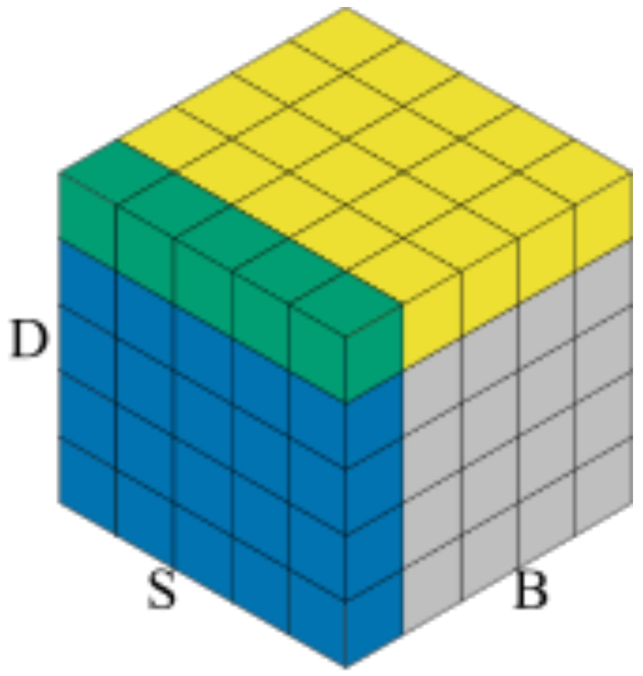
Set norm
Per set standardization
Per feature transformation

Deep Sets ++ and Set Transformer ++

- Careful design of residual connections, **Clean path residual connections**
- Normalization layer specific for sets, **Set Norm**



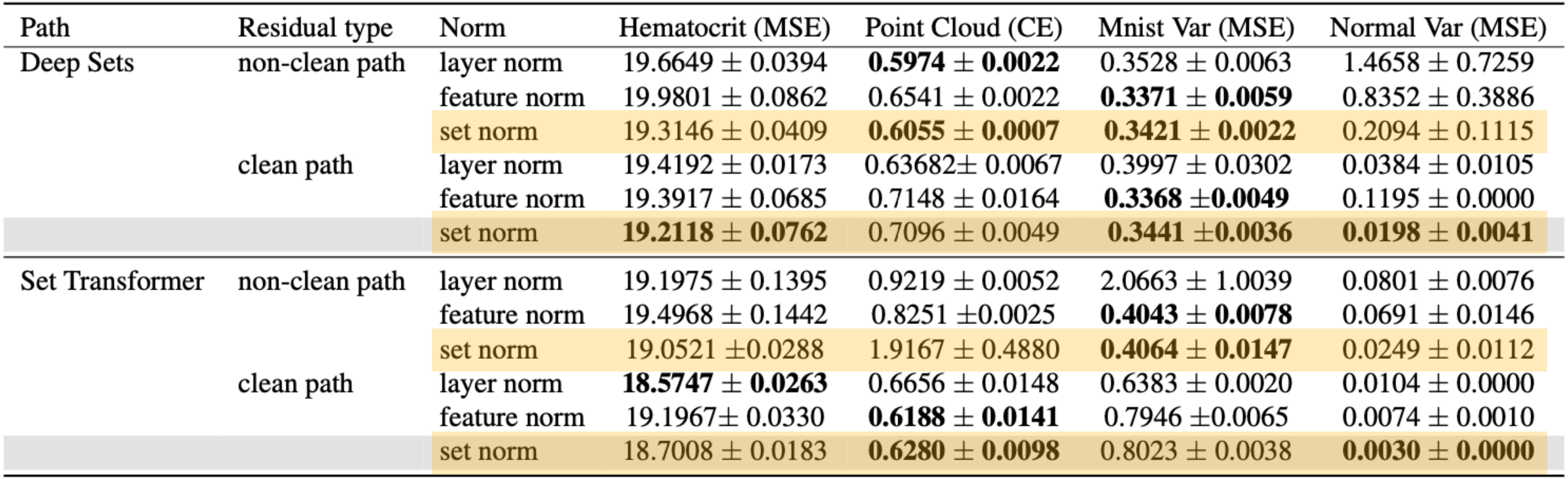
Layer norm
Per set, per sample standardization
Per feature transformation



Set norm
Per set standardization
Per feature transformation

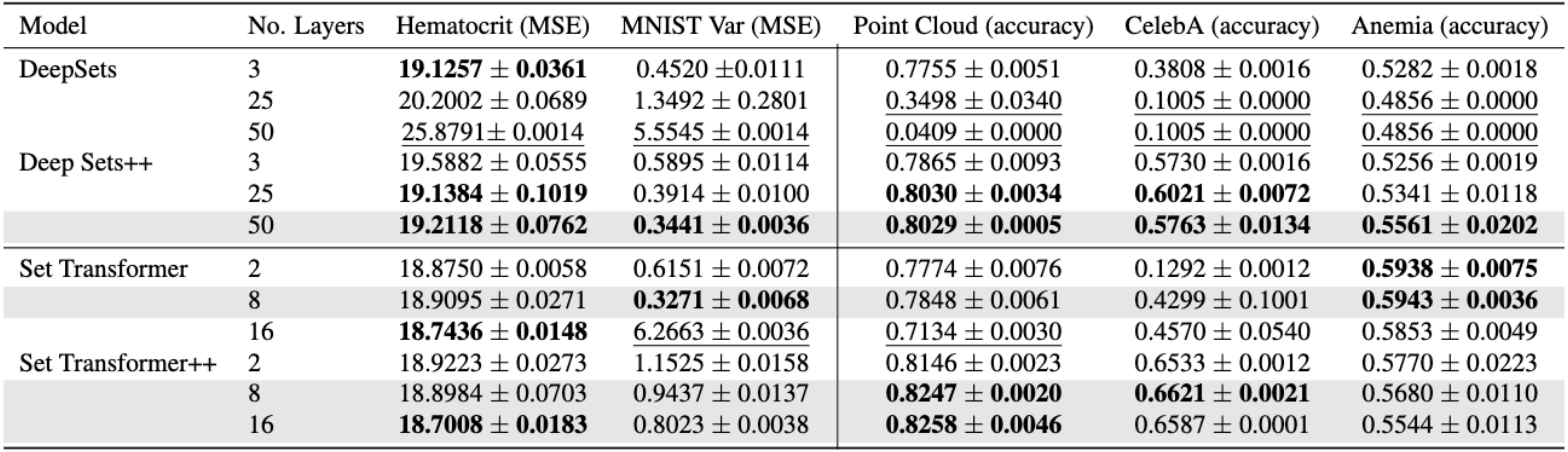
- **Less unrecoverable information**
- **No batch considerations**
- **Permutation equivariant**

Set norm performs better than other norms



Path	Residual type	Norm	Hematocrit (MSE)	Point Cloud (CE)	Mnist Var (MSE)	Normal Var (MSE)
Deep Sets	non-clean path	layer norm	19.6649 ± 0.0394	**0.5974 ± 0.0022**	0.3528 ± 0.0063	1.4658 ± 0.7259
		feature norm	19.9801 ± 0.0862	0.6541 ± 0.0022	**0.3371 ± 0.0059**	0.8352 ± 0.3886
		set norm	19.3146 ± 0.0409	**0.6055 ± 0.0007**	**0.3421 ± 0.0022**	0.2094 ± 0.1115
	clean path	layer norm	19.4192 ± 0.0173	0.63682± 0.0067	0.3997 ± 0.0302	0.0384 ± 0.0105
		feature norm	19.3917 ± 0.0685	0.7148 ± 0.0164	**0.3368 ±0.0049**	0.1195 ± 0.0000
		set norm	**19.2118 ± 0.0762**	0.7096 ± 0.0049	**0.3441 ±0.0036**	**0.0198 ± 0.0041**
Set Transformer	non-clean path	layer norm	19.1975 ± 0.1395	0.9219 ± 0.0052	2.0663 ± 1.0039	0.0801 ± 0.0076
		feature norm	19.4968 ± 0.1442	0.8251 ±0.0025	**0.4043 ± 0.0078**	0.0691 ± 0.0146
		set norm	19.0521 ±0.0288	1.9167 ± 0.4880	**0.4064 ± 0.0147**	0.0249 ± 0.0112
	clean path	layer norm	**18.5747 ± 0.0263**	0.6656 ± 0.0148	0.6383 ± 0.0020	0.0104 ± 0.0000
		feature norm	19.1967± 0.0330	**0.6188 ± 0.0141**	0.7946 ±0.0065	0.0074 ± 0.0010
		set norm	18.7008 ± 0.0183	**0.6280 ± 0.0098**	0.8023 ± 0.0038	**0.0030 ± 0.0000**

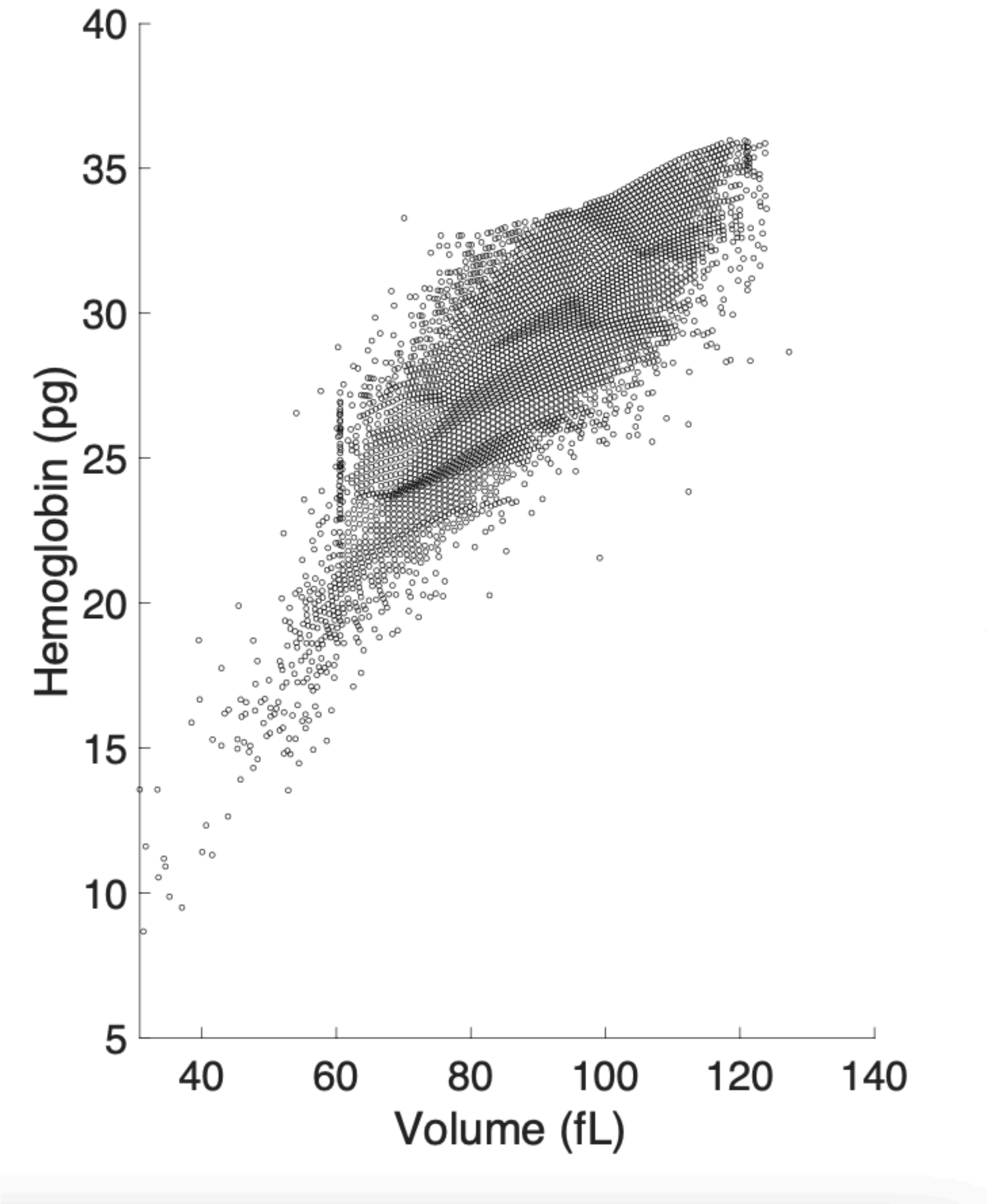
Deep Sets++ and Set Transformer++ reach high depth with improved performances

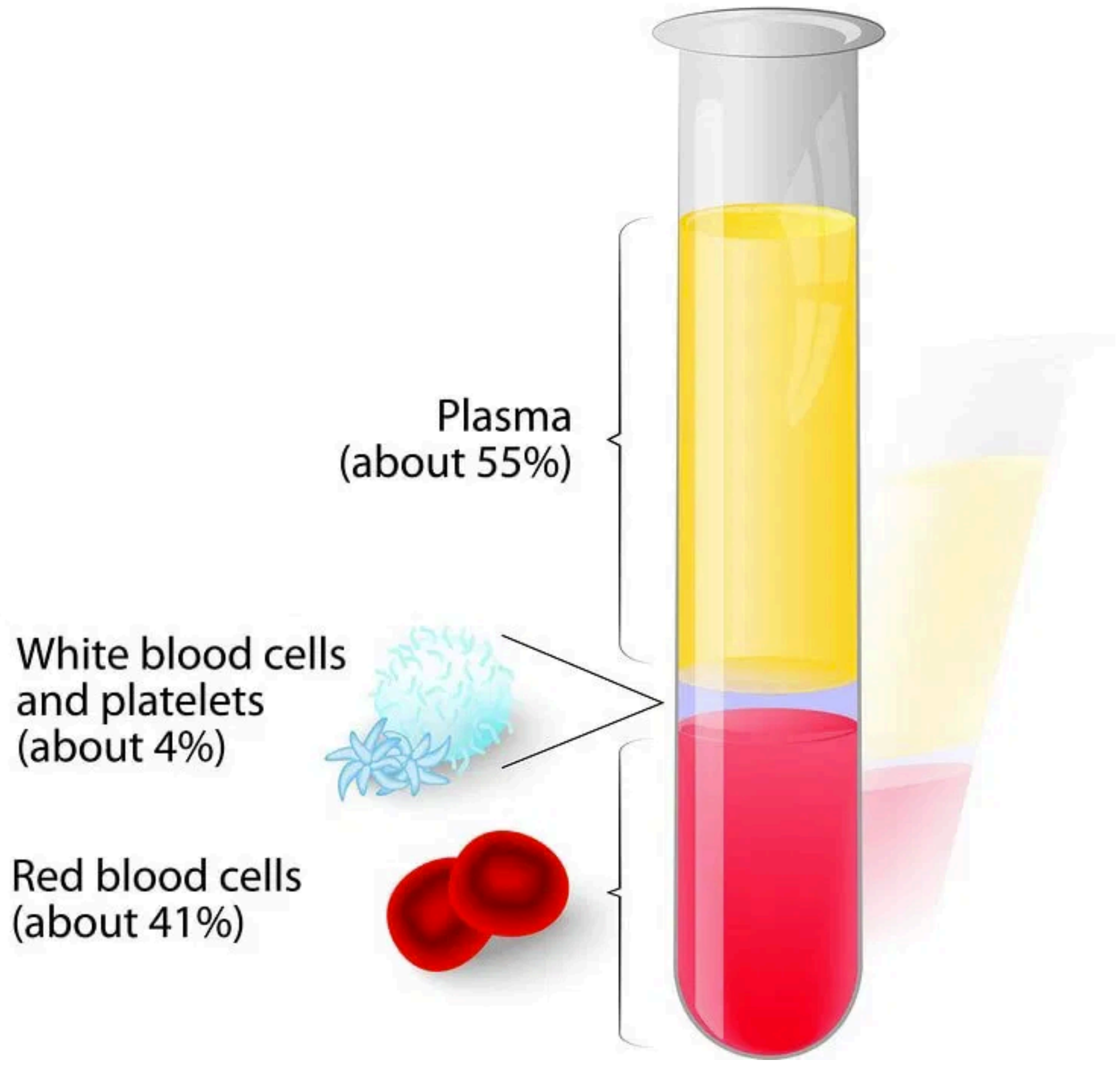


Model	No. Layers	Hematocrit (MSE)	MNIST Var (MSE)	Point Cloud (accuracy)	CelebA (accuracy)	Anemia (accuracy)
DeepSets	3	**19.1257 ± 0.0361**	0.4520 ±0.0111	0.7755 ± 0.0051	0.3808 ± 0.0016	0.5282 ± 0.0018
	25	20.2002 ± 0.0689	1.3492 ± 0.2801	0.3498 ± 0.0340	0.1005 ± 0.0000	0.4856 ± 0.0000
	50	25.8791± 0.0014	5.5545 ± 0.0014	0.0409 ± 0.0000	0.1005 ± 0.0000	0.4856 ± 0.0000
Deep Sets++	3	19.5882 ± 0.0555	0.5895 ± 0.0114	0.7865 ± 0.0093	0.5730 ± 0.0016	0.5256 ± 0.0019
	25	**19.1384 ± 0.1019**	0.3914 ± 0.0100	**0.8030 ± 0.0034**	**0.6021 ± 0.0072**	0.5341 ± 0.0118
	50	**19.2118 ± 0.0762**	**0.3441 ± 0.0036**	**0.8029 ± 0.0005**	**0.5763 ± 0.0134**	**0.5561 ± 0.0202**
Set Transformer	2	18.8750 ± 0.0058	0.6151 ± 0.0072	0.7774 ± 0.0076	0.1292 ± 0.0012	**0.5938 ± 0.0075**
	8	18.9095 ± 0.0271	**0.3271 ± 0.0068**	0.7848 ± 0.0061	0.4299 ± 0.1001	**0.5943 ± 0.0036**
	16	**18.7436 ± 0.0148**	6.2663 ± 0.0036	0.7134 ± 0.0030	0.4570 ± 0.0540	0.5853 ± 0.0049
Set Transformer++	2	18.9223 ± 0.0273	1.1525 ± 0.0158	0.8146 ± 0.0023	0.6533 ± 0.0012	0.5770 ± 0.0223
	8	18.8984 ± 0.0703	0.9437 ± 0.0137	**0.8247 ± 0.0020**	**0.6621 ± 0.0021**	0.5680 ± 0.0110
	16	**18.7008 ± 0.0183**	0.8023 ± 0.0038	**0.8258 ± 0.0046**	0.6587 ± 0.0001	0.5544 ± 0.0113

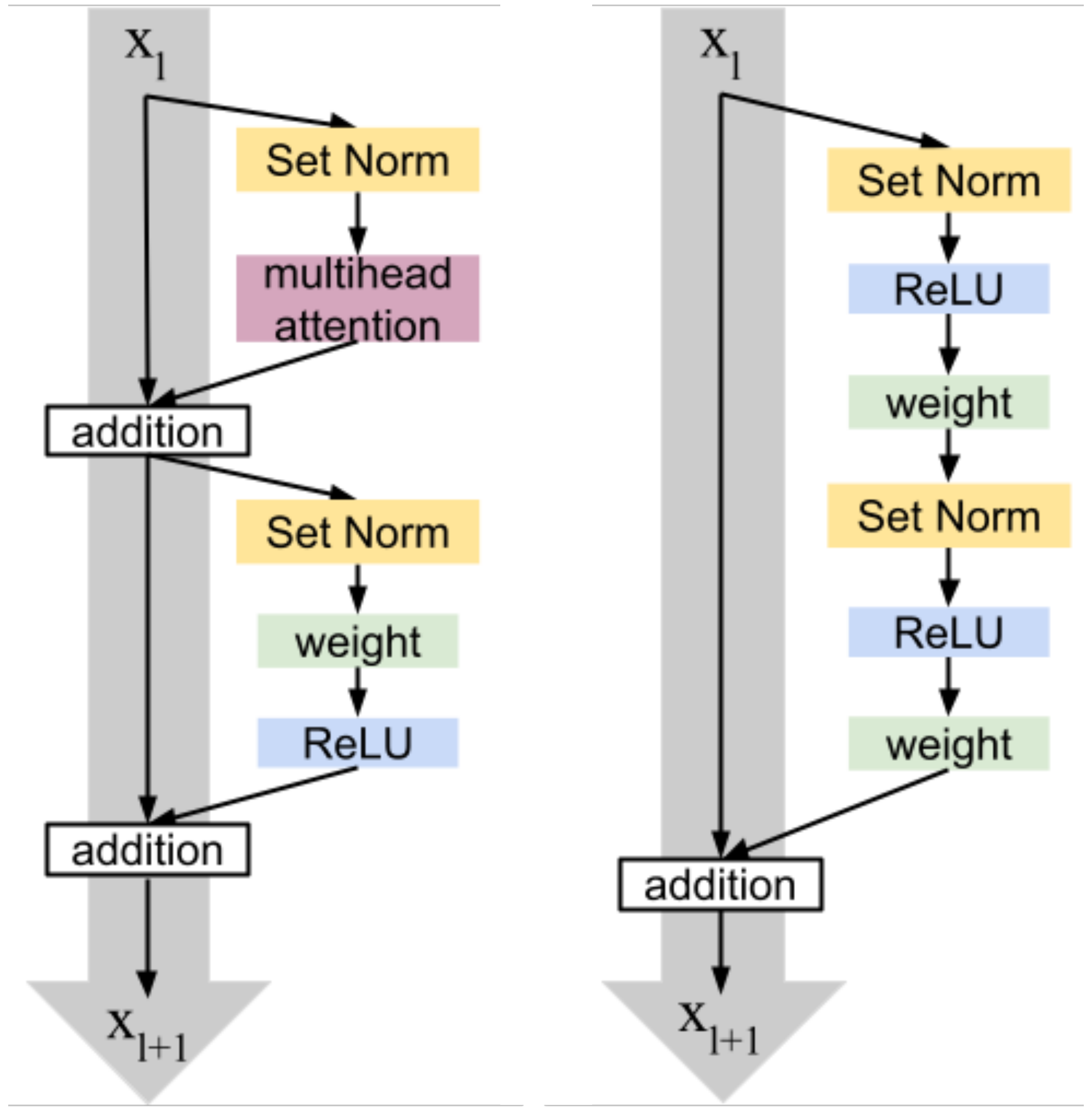
Flow RBC

- >100,000 patients, collected over 6 years
- Input: single-cell measurements
- Output: cell population property (hematocrit)
- Open-sourced for benchmarking!

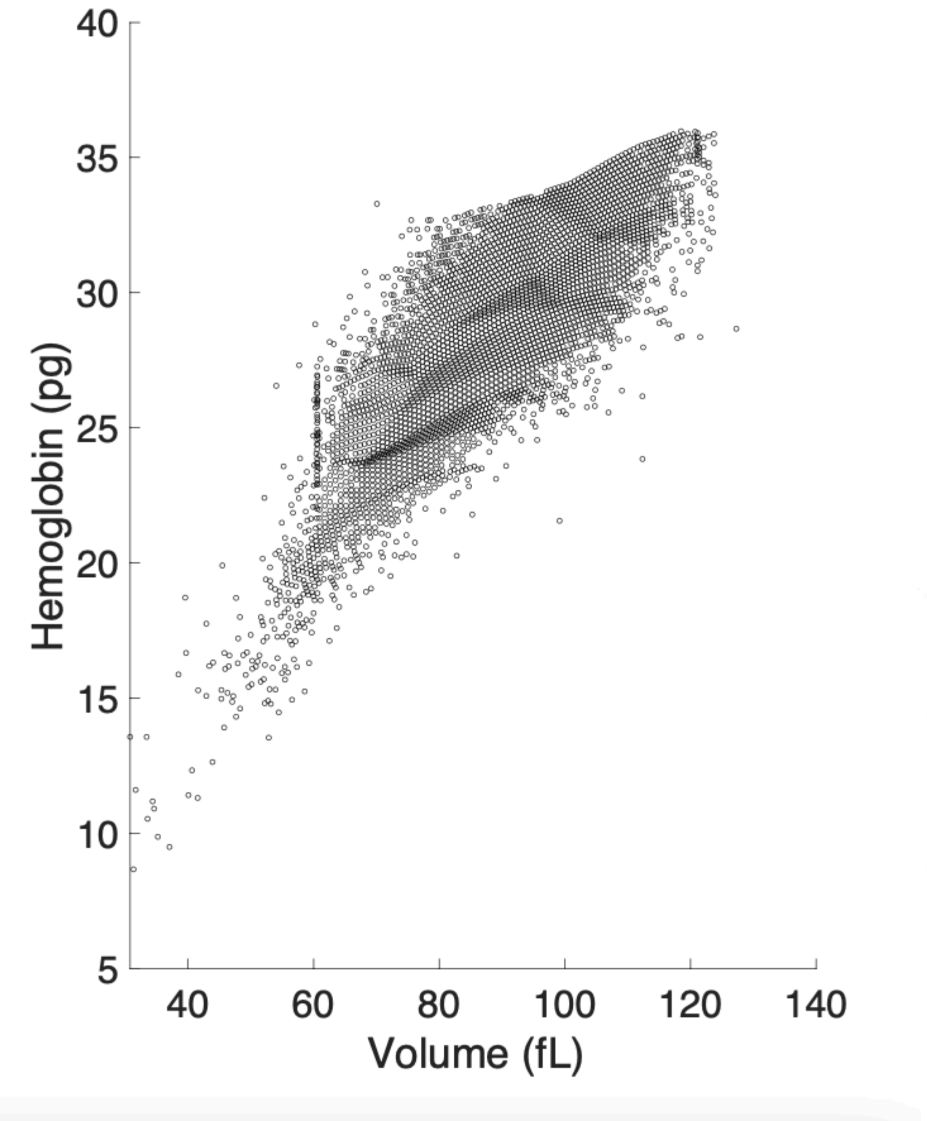


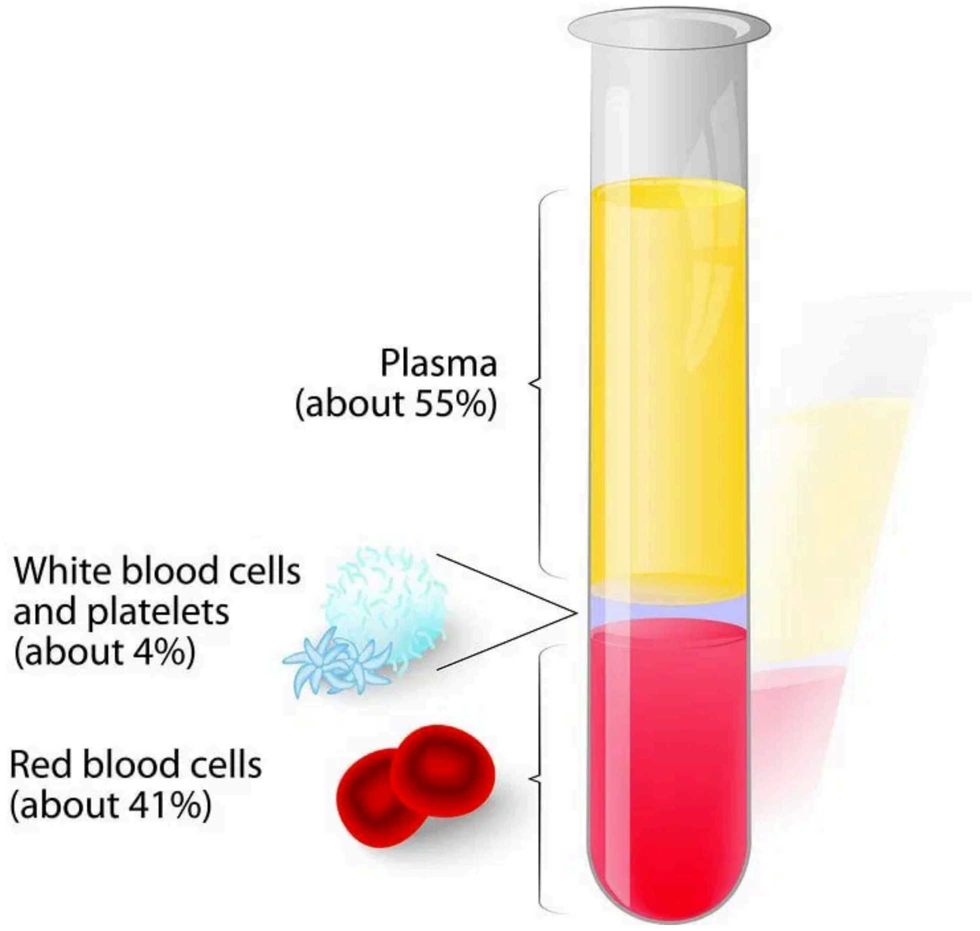


Set Transformer++/ Deep Sets++



Flow RBC





Come visit us at our poster: Hall E #524!