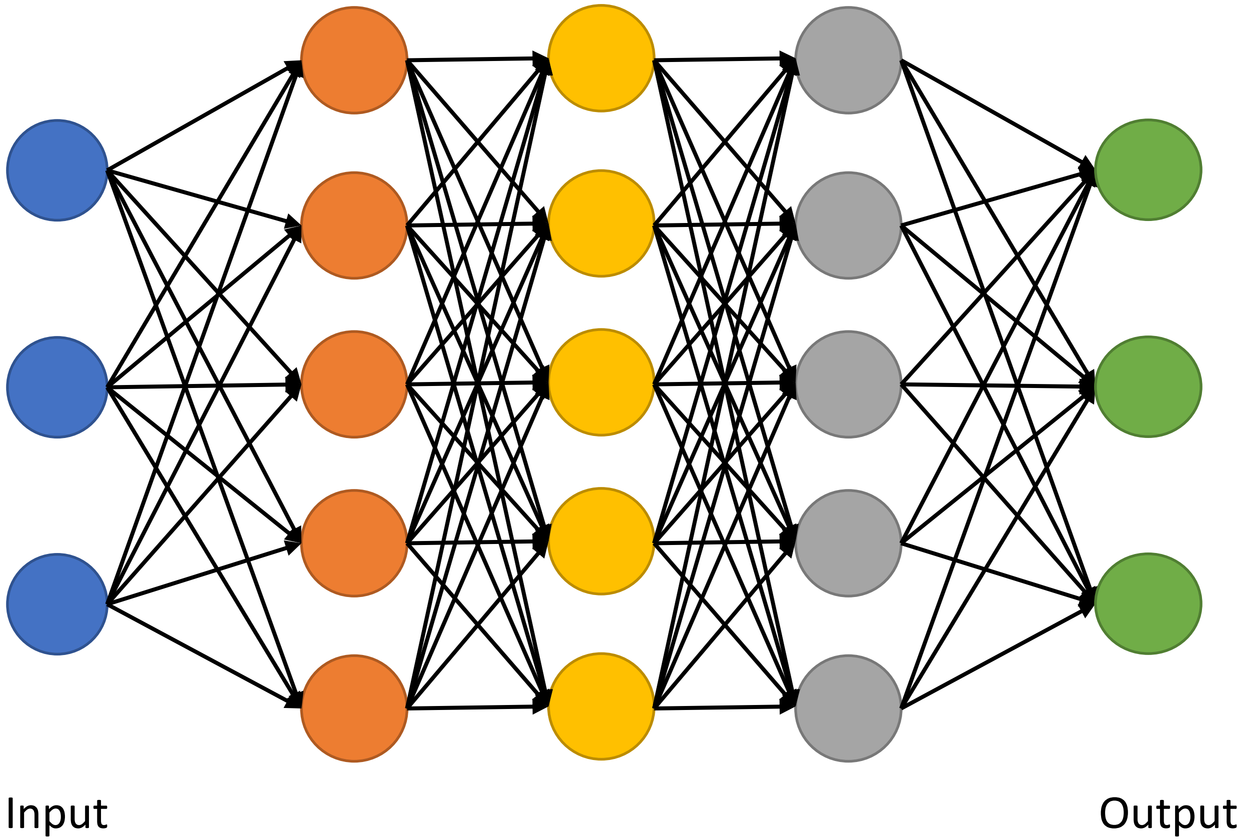
Shapley Explanation Networks

Rui Wang, Xiaoqian (Joy) Wang, David I. Inouye
{ruiw, joywang, dinouye}@purdue.edu
School of Electrical and Computer Engineering
Purdue University
ICLR 2021

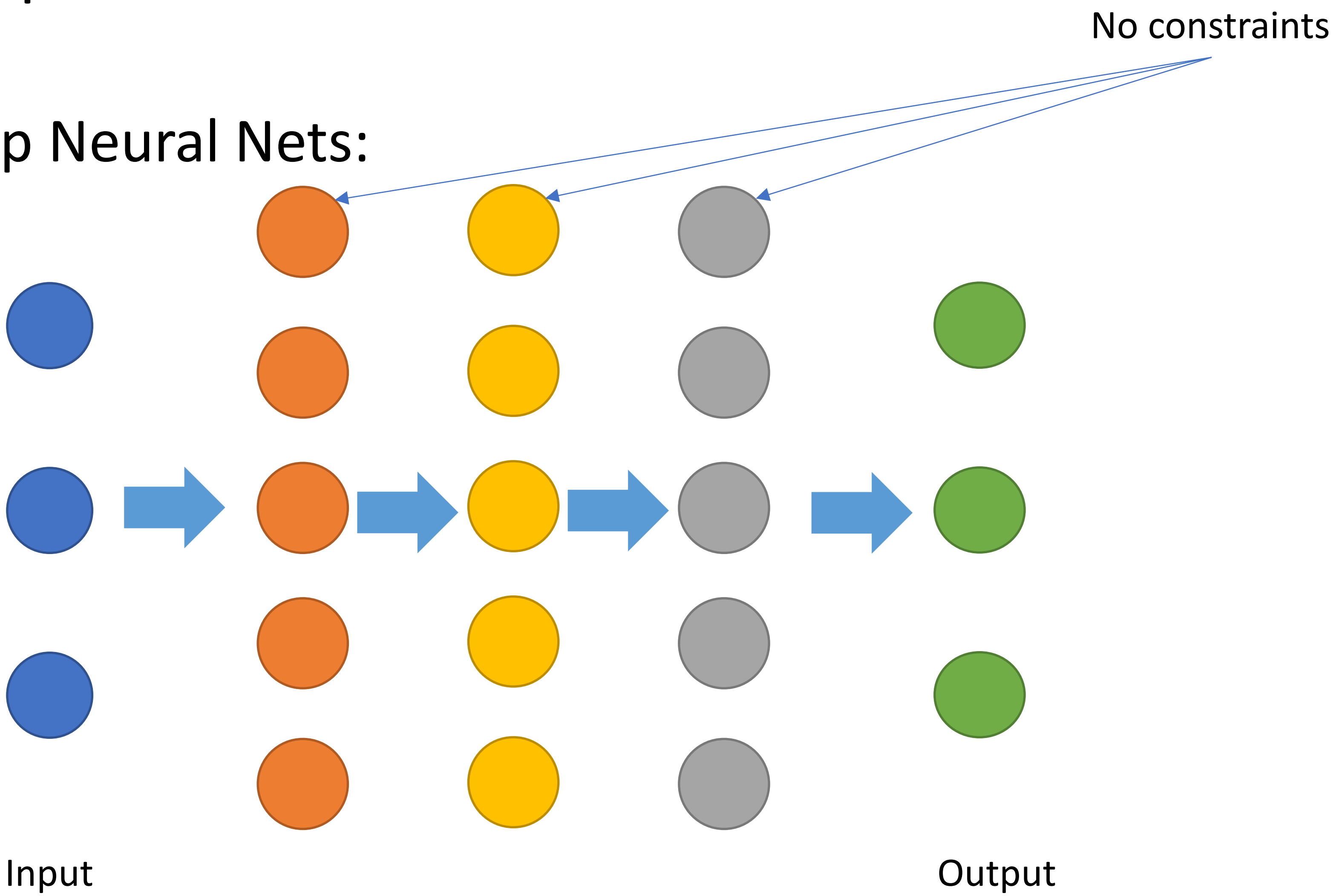
Shapley Explanation Networks

- Canonical Deep Neural Nets:



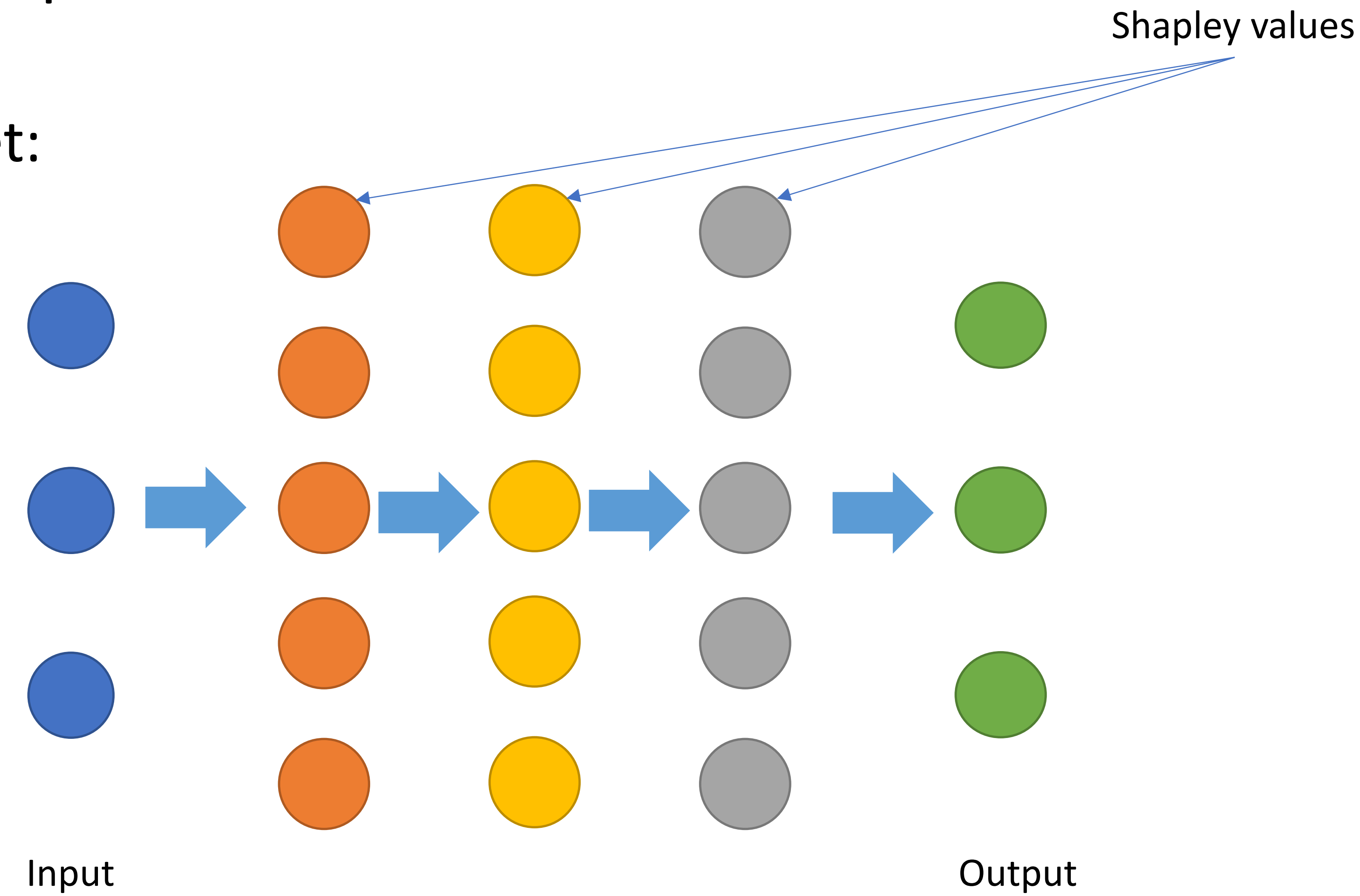
Shapley Explanation Networks

- Canonical Deep Neural Nets:

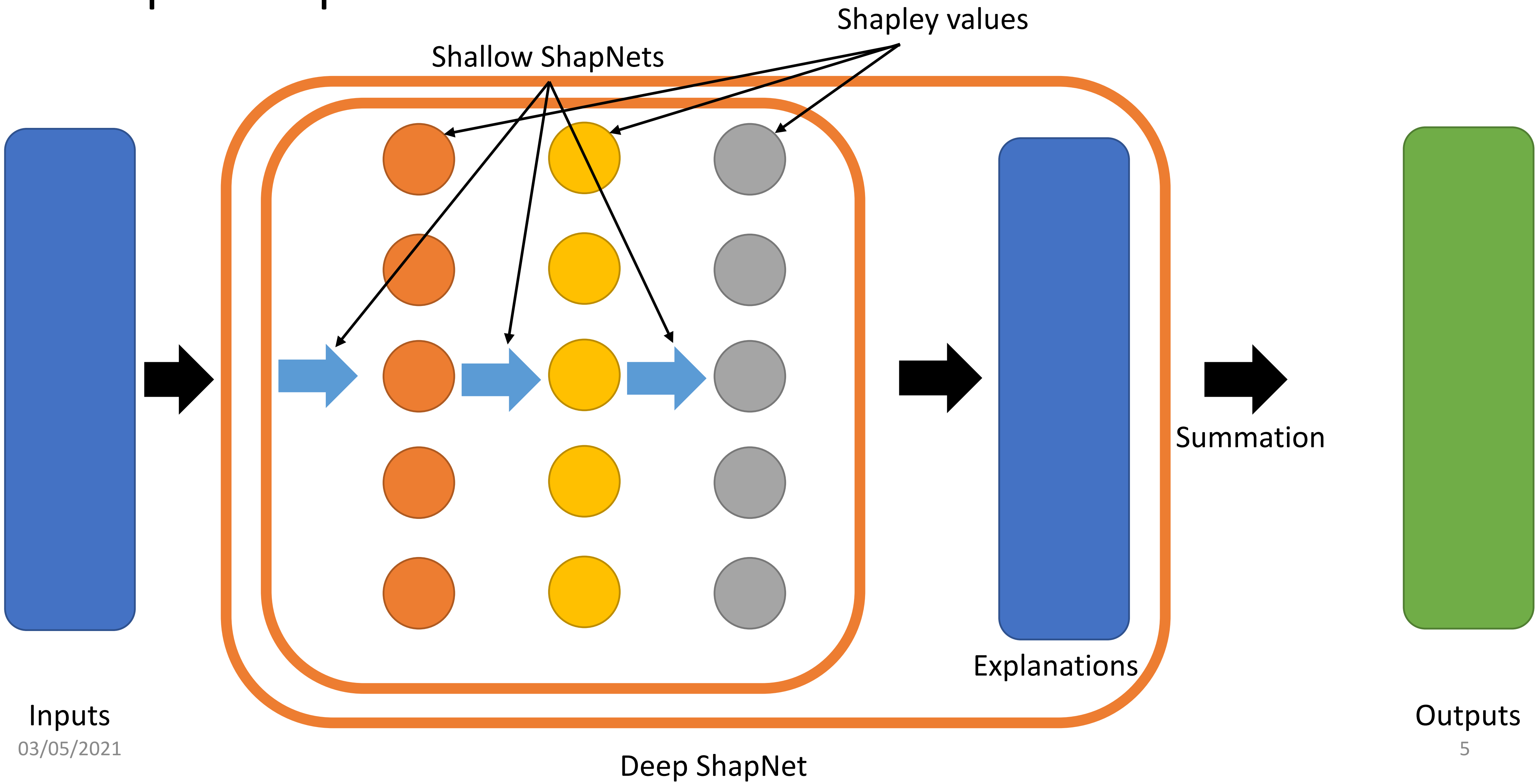


Shapley Explanation Networks

- Deep ShapNet:

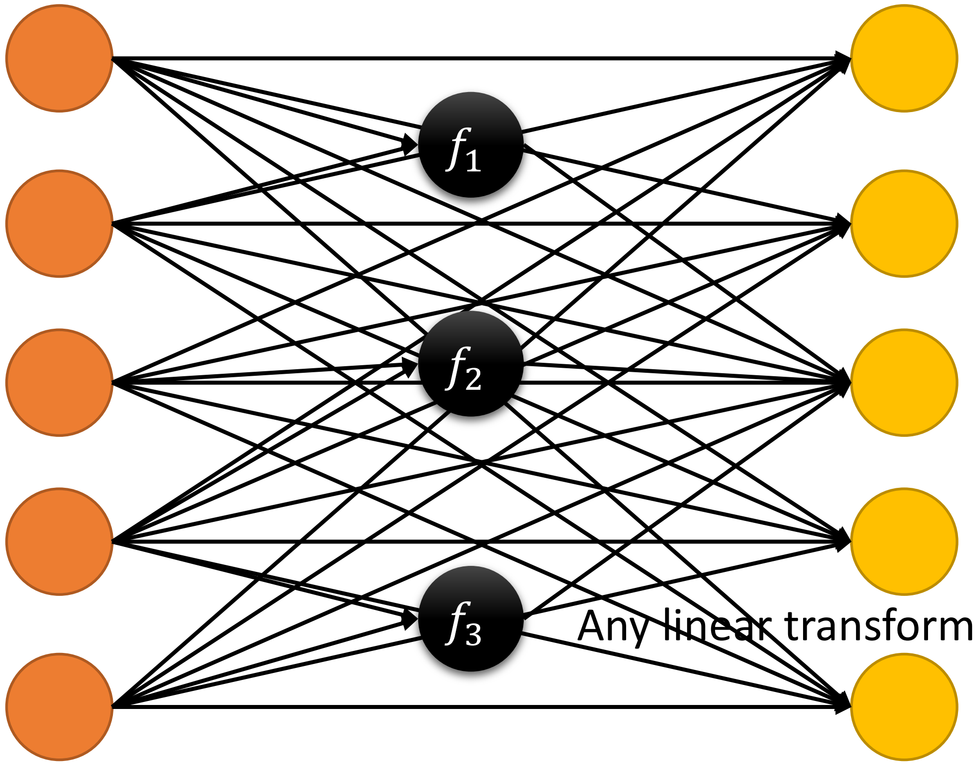


Deep ShapNets



How to overcome the computational burden

- We use super position of small functions to cut down computational requirement



Deep ShapNet for Images

- Notice how each matrix multiplication in convolution is a *small* function
  - Hence we just modify the convolution operator to arrive at convolutional Shapley layer.
- Convolution:
  - Unfold the image representation
  - Matrix multiplication
  - Fold the resulting representation back
- A substitute for pooling:
  - À-trous (dilated) convolution

Experiments: How expressive are ShapNets?

- Not as good, but not bad either

Table 1: Model performance (loss for synthetic or accuracy for others, averaged over 50 runs)

Datasets \ Models	Deep ShapNet	DNN (eq. comp.)	DNN (eq. param.)	Shallow ShapNet	GAM
Synthetic (loss)	3.37e-3	3.93e-3	6.62e-3	3.11e-3	3.36e-3
Yeast	0.585	0.576	0.575	0.577	0.597
Breast Cancer	0.959	0.966	0.971	0.958	0.969

Table 2: Accuracies of Deep ShapNets for images, comparable CNNs and state-of-the-art models.

Datasets \ Models	Deep ShapNet	Comparable CNN	SOTA
MNIST	0.9950	0.9917	0.9984
FashionMNIST	0.9195	0.9168	0.9691
Cifar-10	0.8206	0.7996	0.9970

Experiments: How good are the explanations?

- Pretty good!

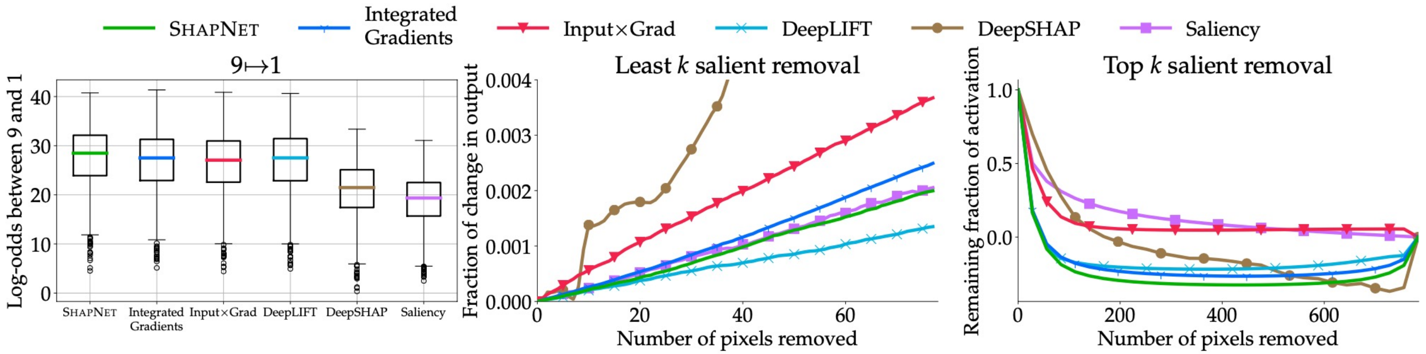


Figure 4: Our intrinsic Deep SHAPNET explanations perform better than post-hoc explanations in identifying the features that can flip the model prediction or that contribute most to the prediction as in figures on the left showing the results of flipping digits as introduced in Shrikumar et al. (2017) and right showing the remaining activation after removing the top k features identified by each explanation method. While Deep SHAPNET explanations did not perform the best in the middle where we show the results after removing least k-salient features as introduced in Srinivas & Fleuret (2019), our model still scores the second. All results are measured on MNIST test set. More results for digit flipping, in Fig. 11, show the same conclusion with statistical significance in Table 7.

Shapley-based Explanation Regularization

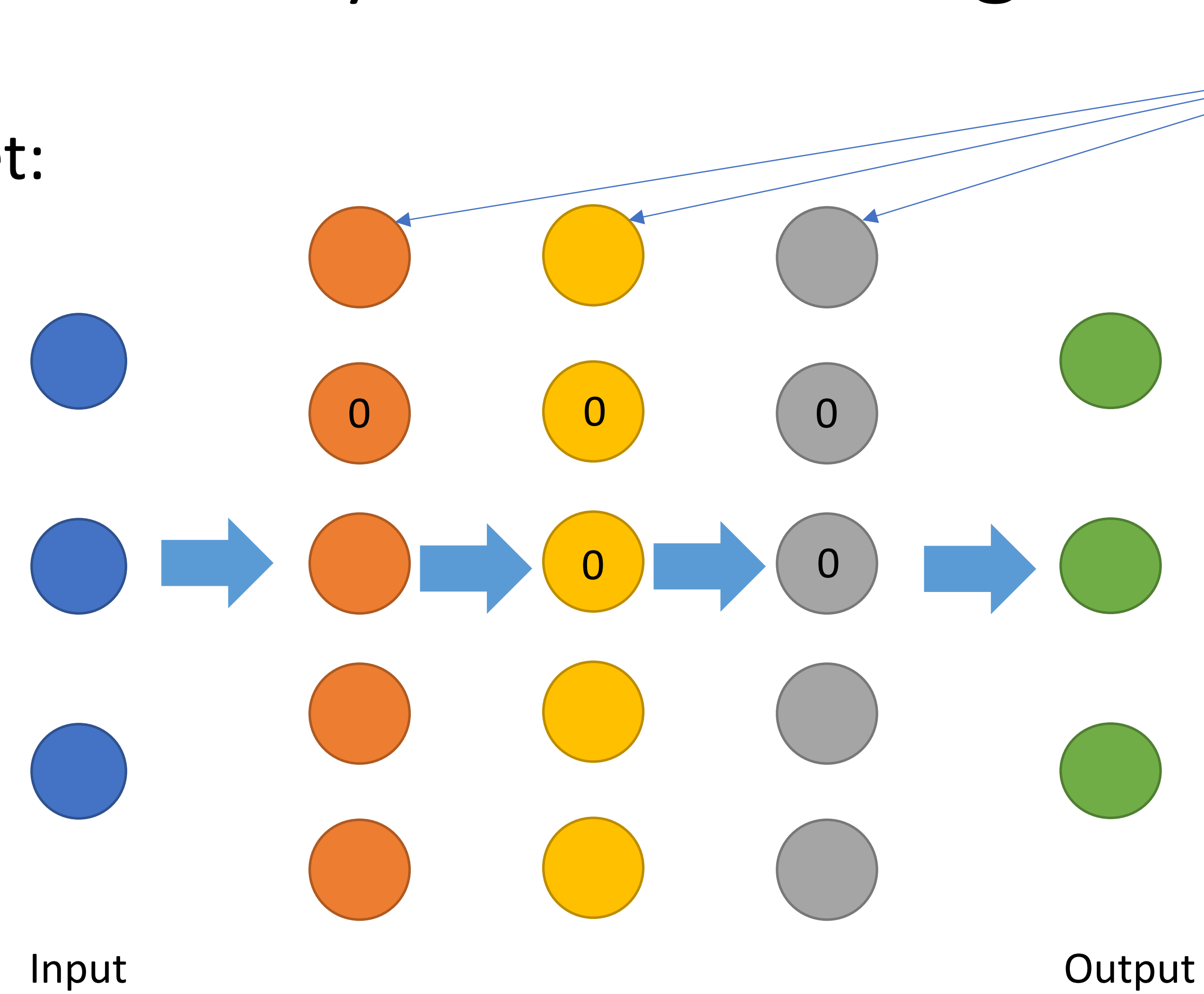
Table 4: Explanation regularization experiments with Deep SHAPNETs (averaged over 50 runs)

Metrics \ Models	Yeast			Breast Cancer Wisconsin		
	ℓ_∞ Reg.	ℓ_1 Reg.	No Reg.	ℓ_∞ Reg.	ℓ_1 Reg.	No Reg.
Coefficient of variation for abs. SHAP	**0.768**	1.23	1.05	**1.28**	NaN	2.04
Sparsity of SHAP values	0.003	**0.00425**	0.00275	0.429	**0.841**	0.209
Accuracy	**0.592**	**0.592**	0.587	0.957	**0.960**	**0.960**

Figure 6: MNIST SHAPNET explanations for different regularizations qualitatively demonstrate the effects of regularization. We notice that ℓ_1 only puts importance on a few key features of the digits while ℓ_∞ spreads out the contribution over more of the image. Red and blue correspond to positive and negative contribution respectively. More visualization of the explanations, including the other classes and more in-depth discussion, can be found in subsection H.4.

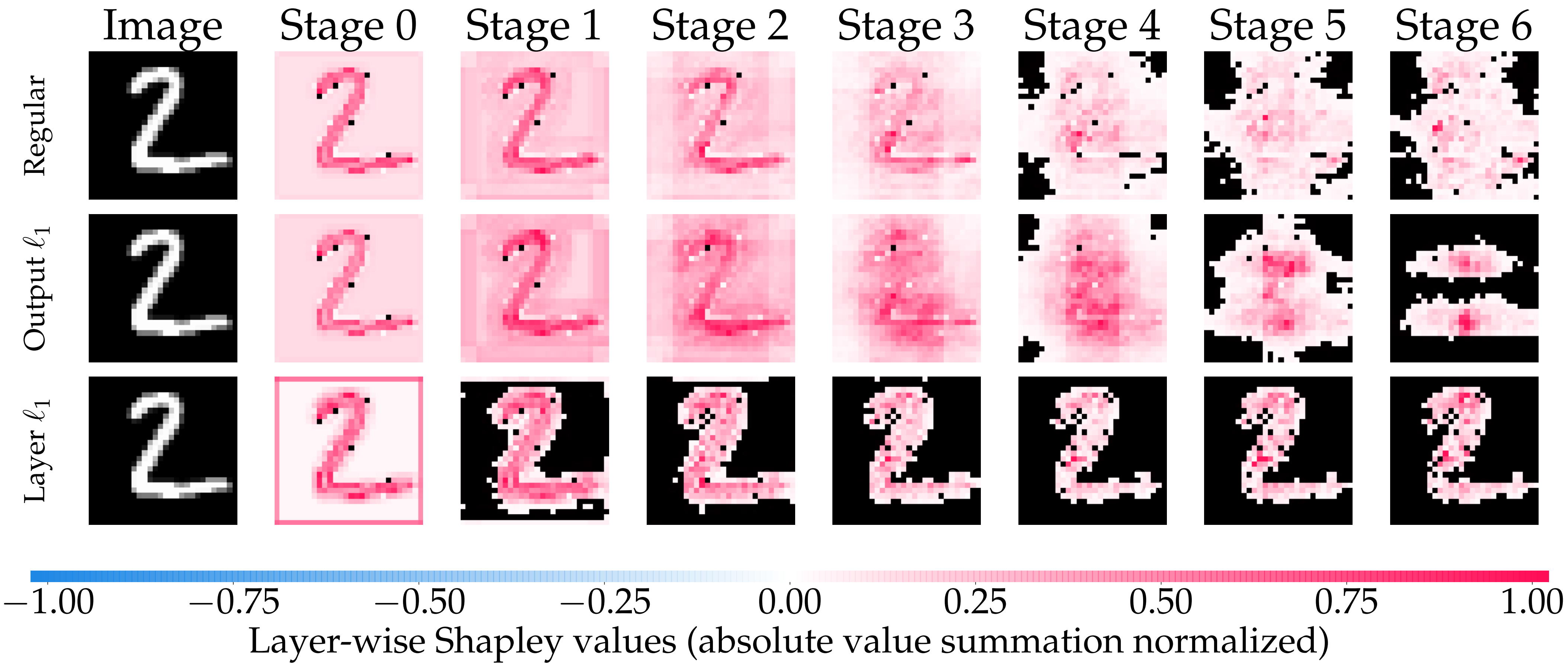
Instance-based Dynamic Pruning

• Deep ShapNet:



By missingness!
(Lundberg & Lee, 2017)

Pruning in action



References & Acknowledgement

- S. Lundberg, S. Lee. A unified approach to interpreting model predictions. NeurIPS, 2017.

- R.W. and D.I. acknowledge support from Northrop Grumman Corporation through the Northrop Grumman Cybersecurity Research Consortium (NGCRC) and the Army Research Lab through contract number W911NF-2020-221. X.W. acknowledges support from NSF III #1955890.