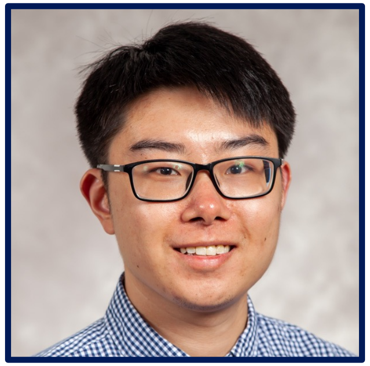
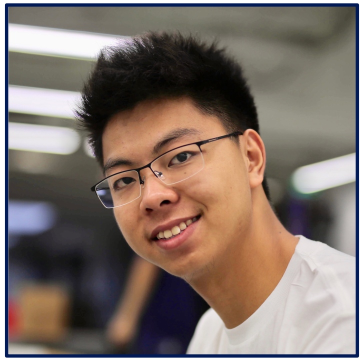
Concept Whitening for Interpretable Image Recognition

Interpretable ML Lab, Duke University



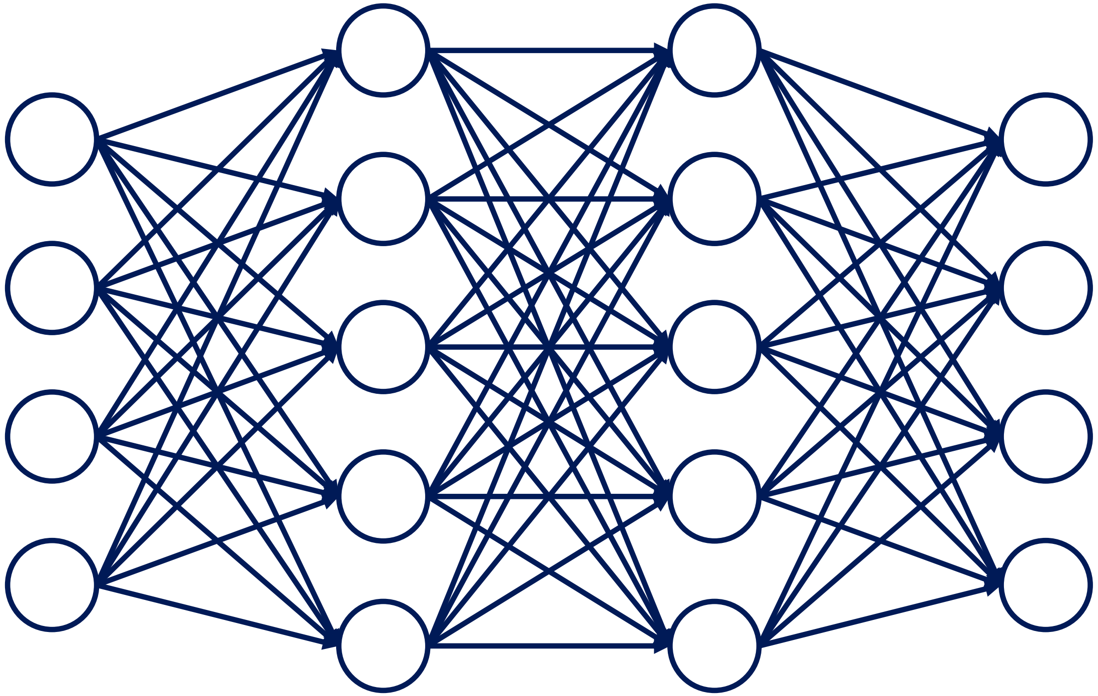
Zhi Chen



Yijie Bei

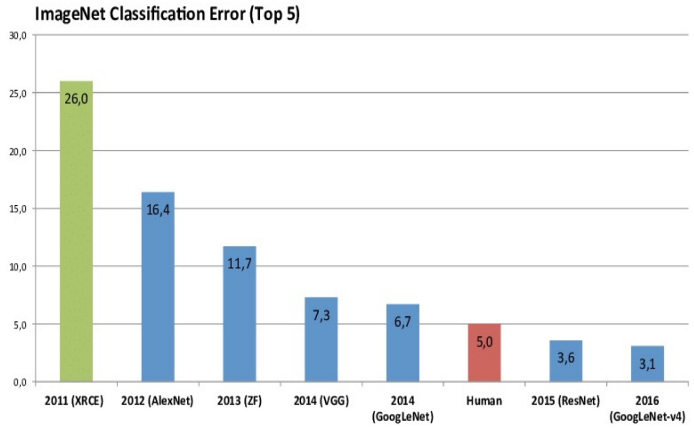
Cynthia Rudin

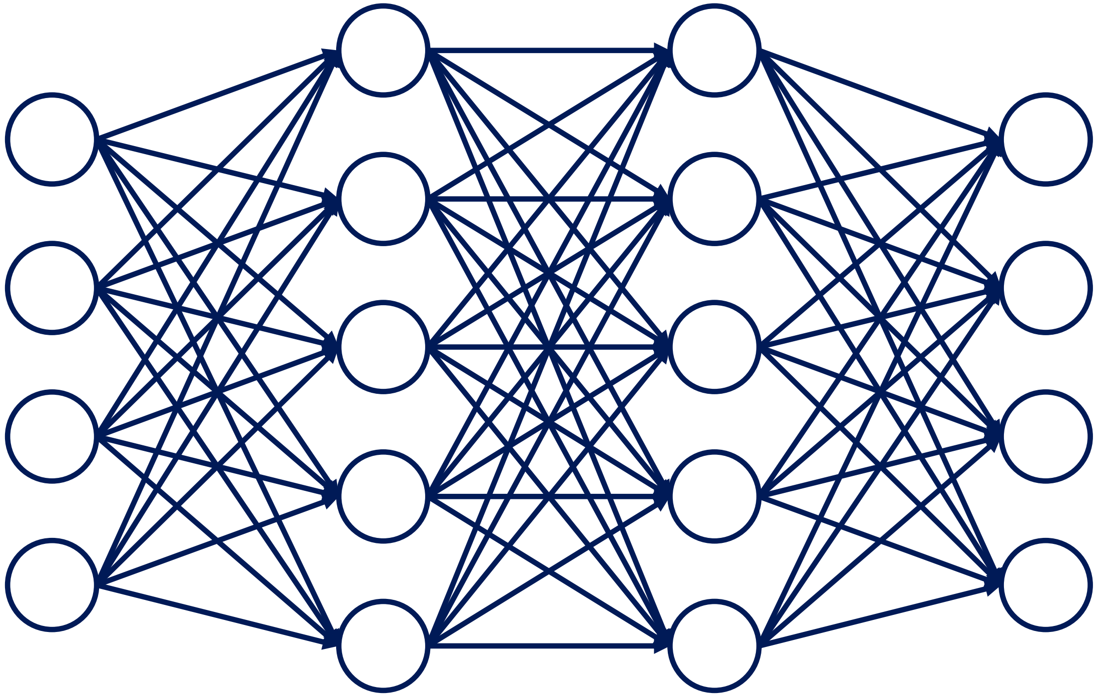
Duke

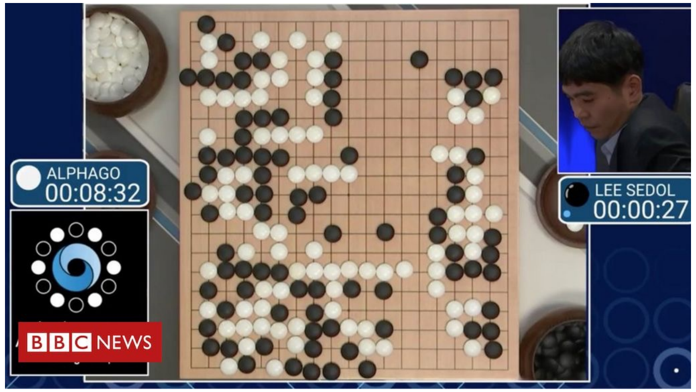


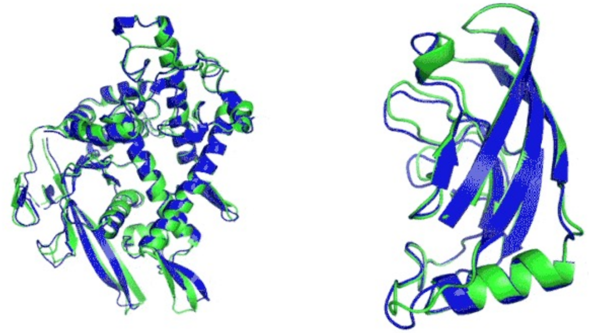
Duke

Great success !!!



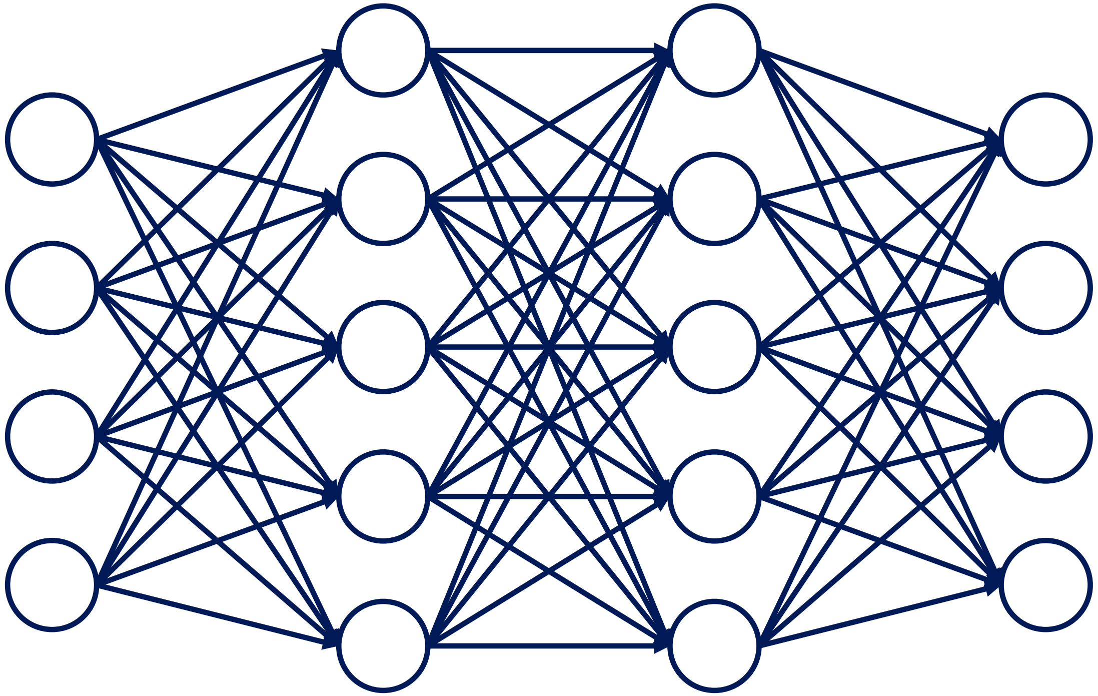






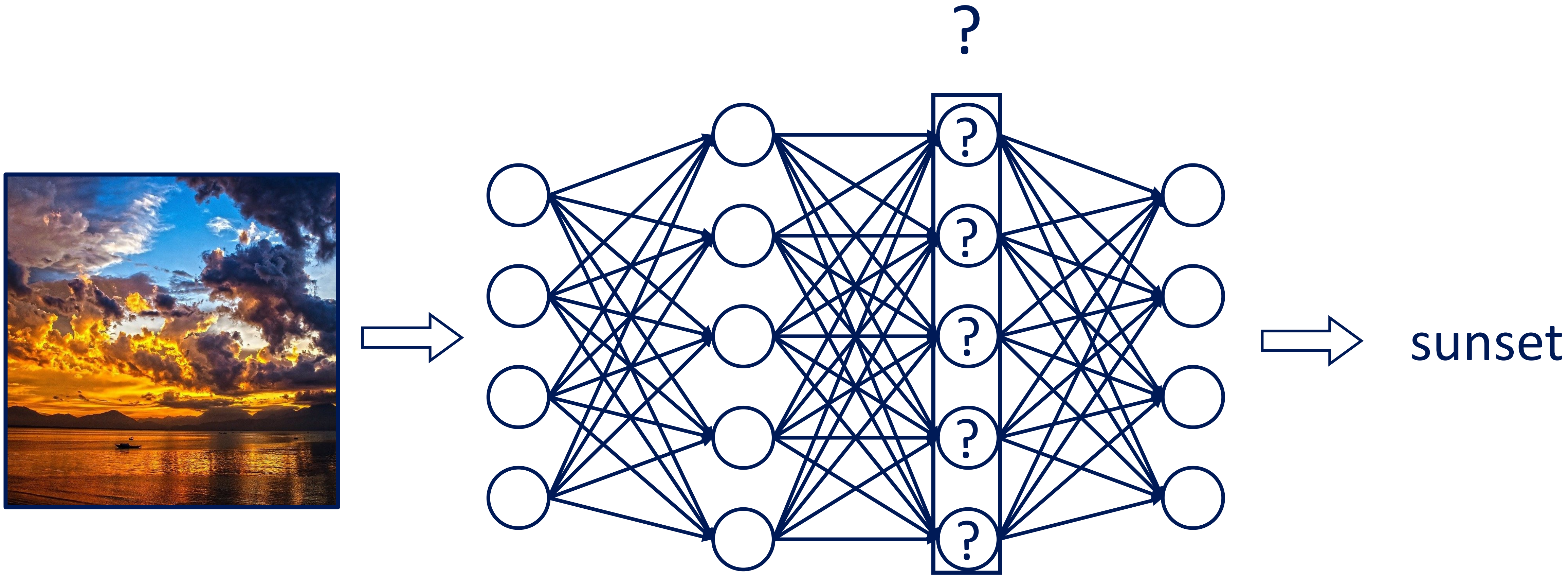
T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

T1049 / 6y4f
93.3 GDT
(adhesin tip)



sunset

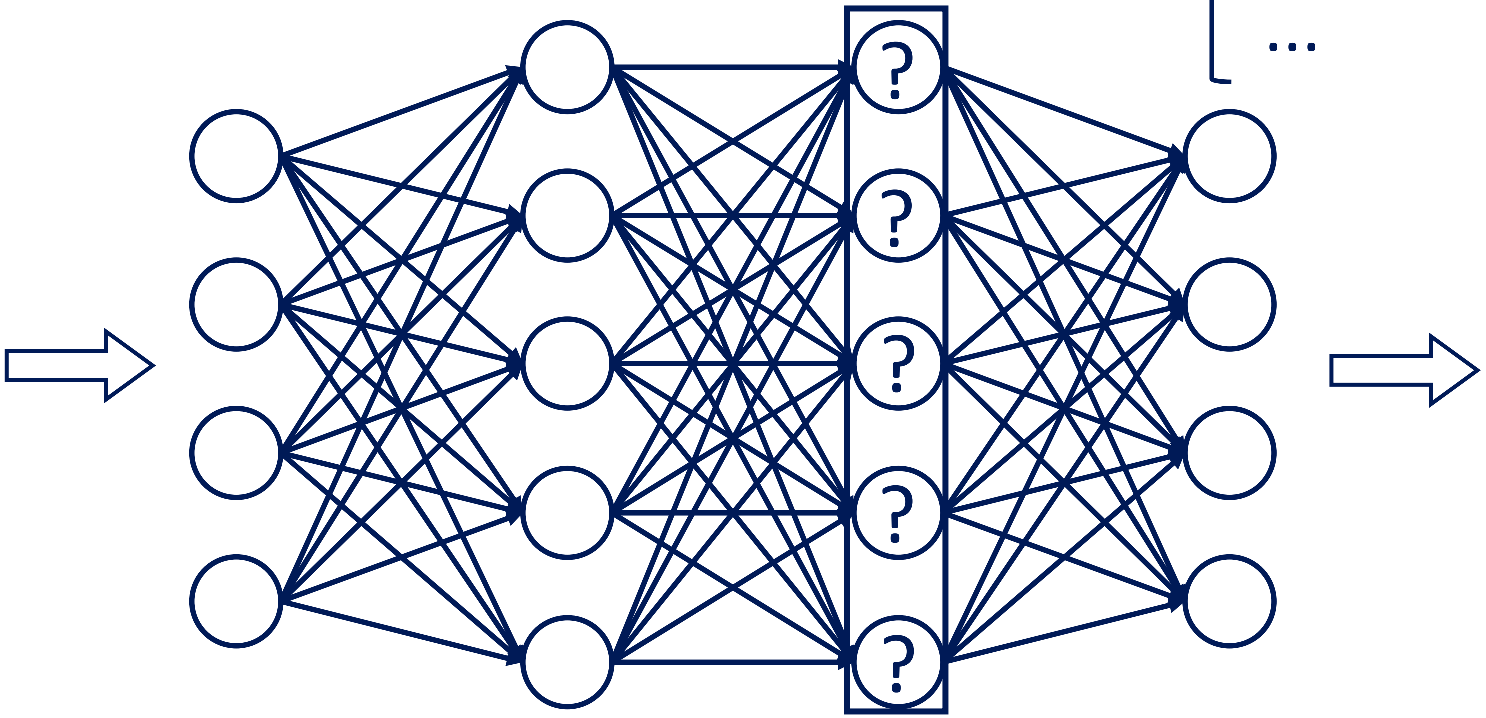
Duke


?
?
?
?
?
?
sunset

interpretability

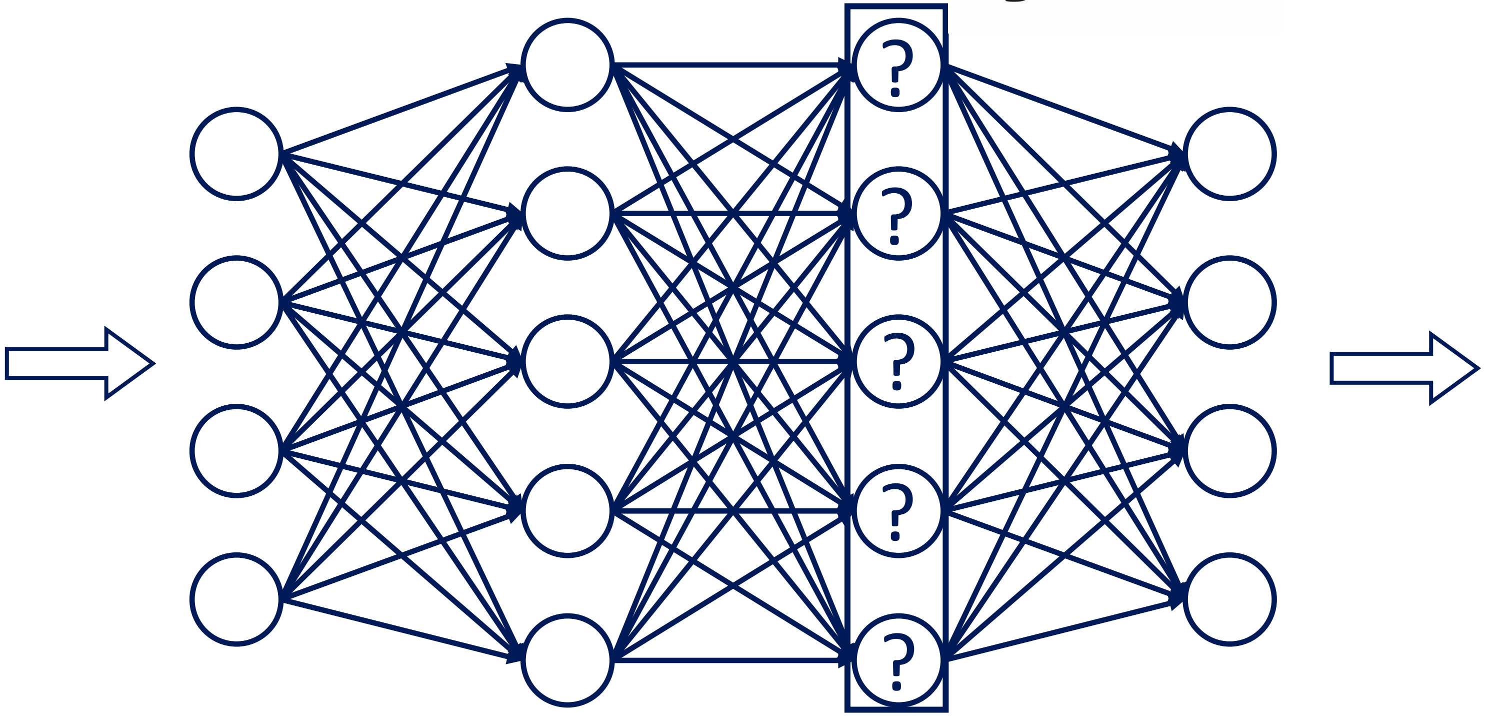
1. build trust
2. debug models
3. gain knowledge

...



sunset

post hoc analysis



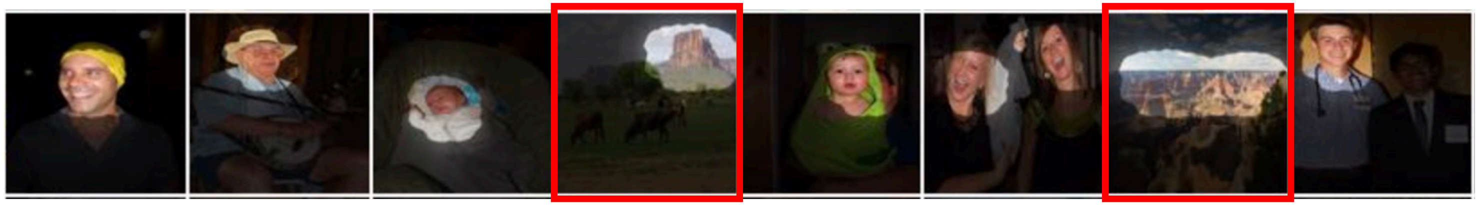
sunset

concept based - human reason in concepts

Duke

Post hoc analysis – concept based

- Single neuron (Zhou et al, 2014; 2018)

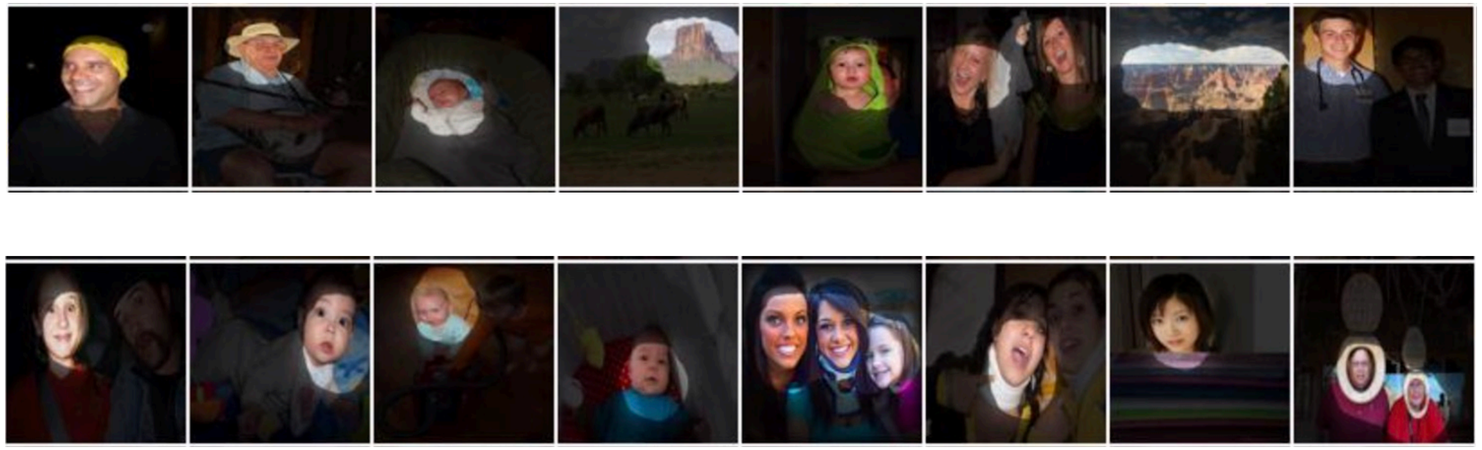


neuron 1

impure!

Post hoc analysis – concept based

- Single neuron (Zhou et al, 2014; 2018)



neuron 1

neuron 2

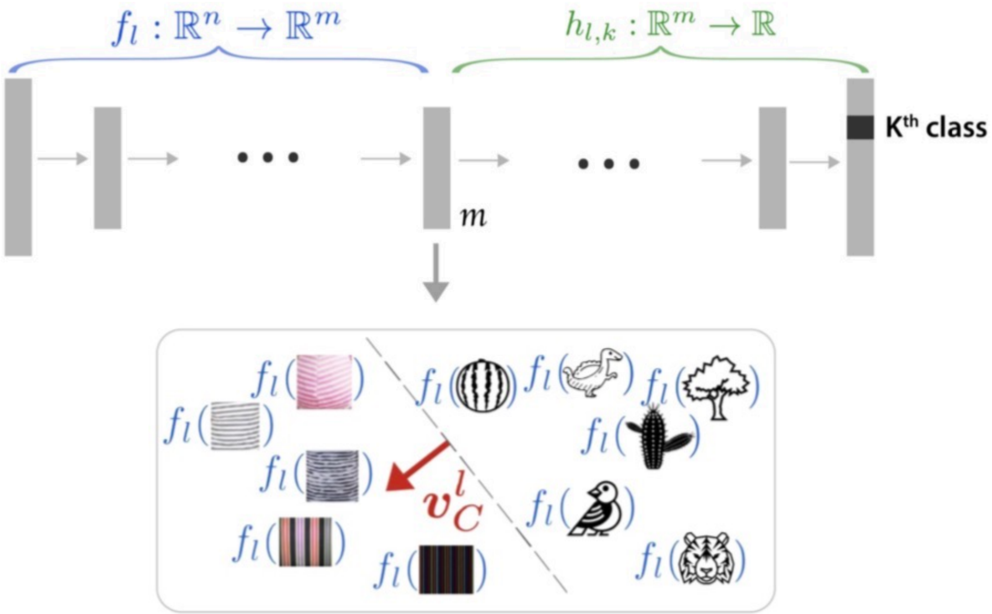
distributed!

single neuron of standard NN --> single concept

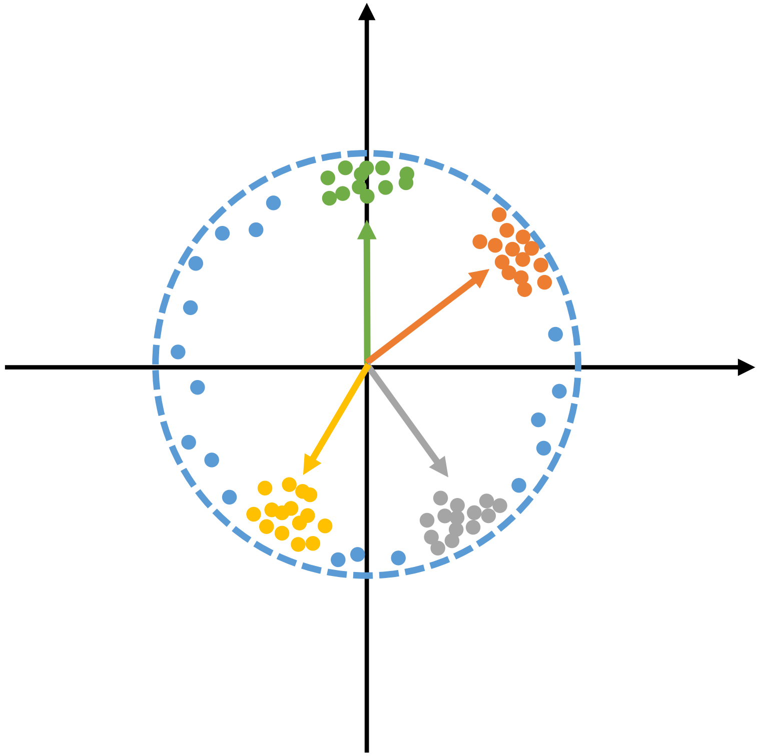
Duke

Post hoc analysis – concept based

- Linear combination of neurons (Kim et al, 2017; Zhou et al, 2018)
  - better than single neuron
  - assume all concepts are separated



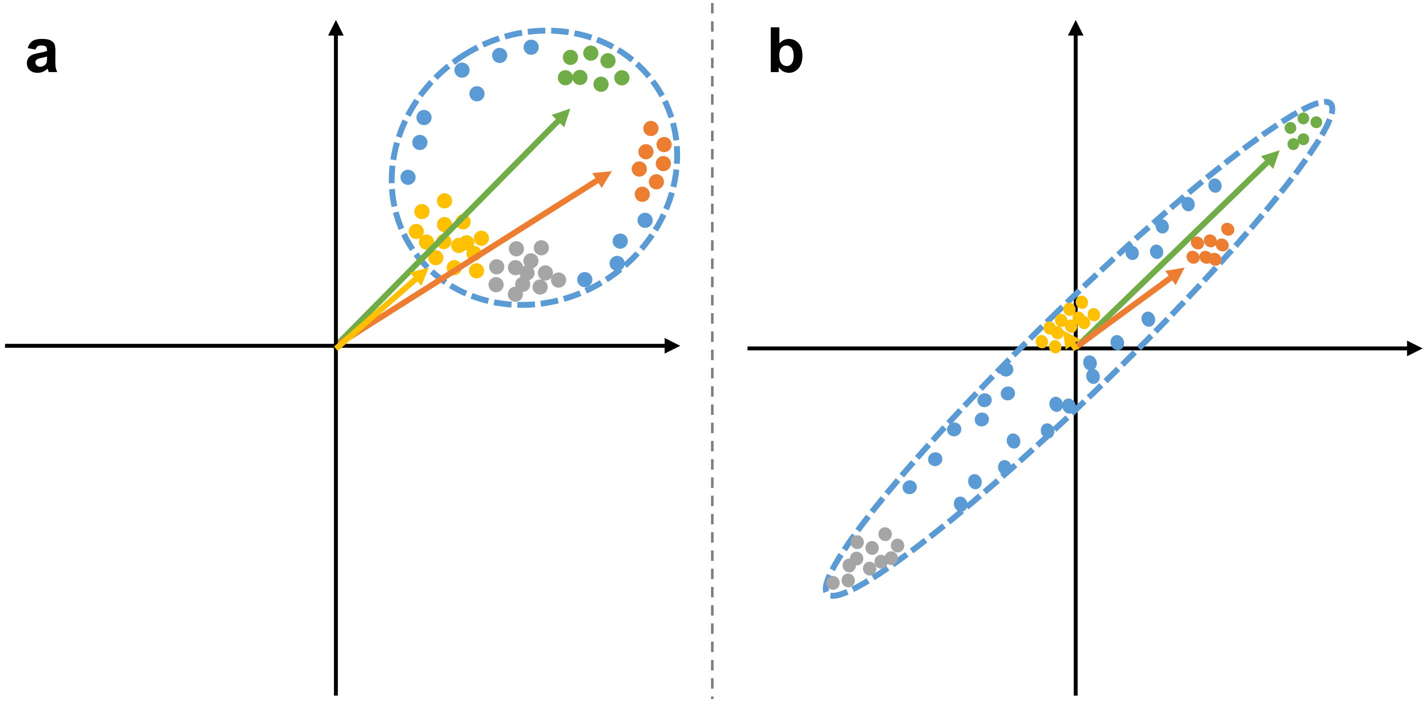
TCAV, Kim et al, 2017



Duke

Post hoc analysis – concept based

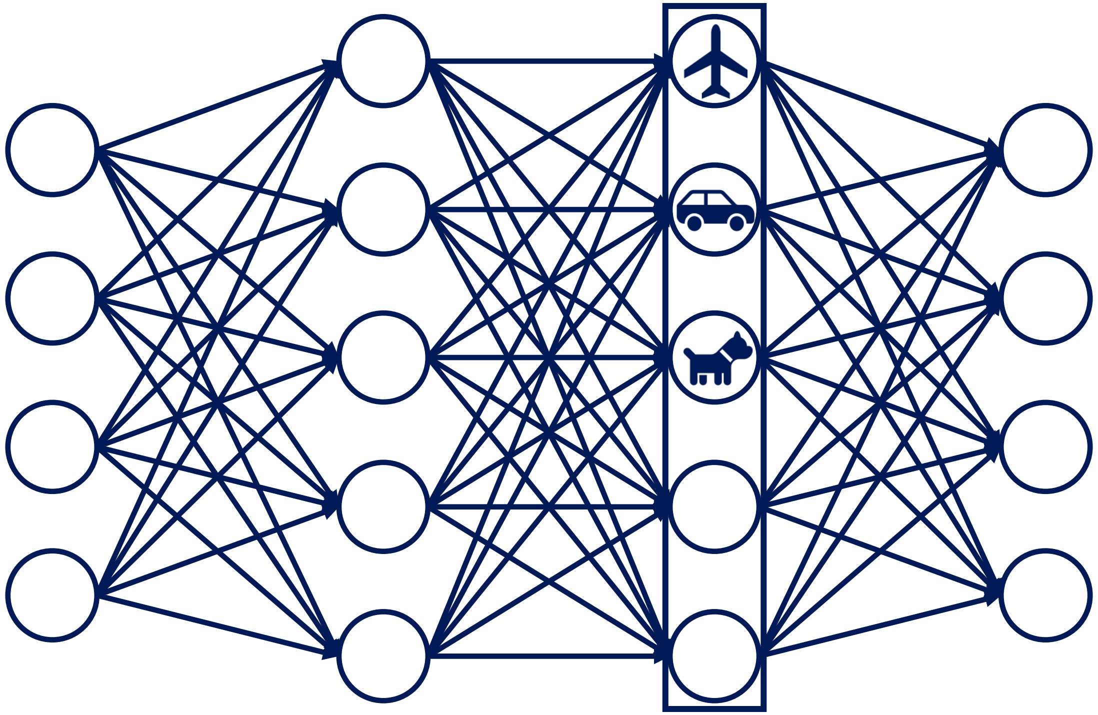
- Linear combination of neurons (Kim et al, 2017; Zhou et al, 2018)
  - reality: concept vectors may point to the same direction



Duke

The idea

- Why not do it by ourselves?
- Create a disentangled latent space that its axes represent known concepts

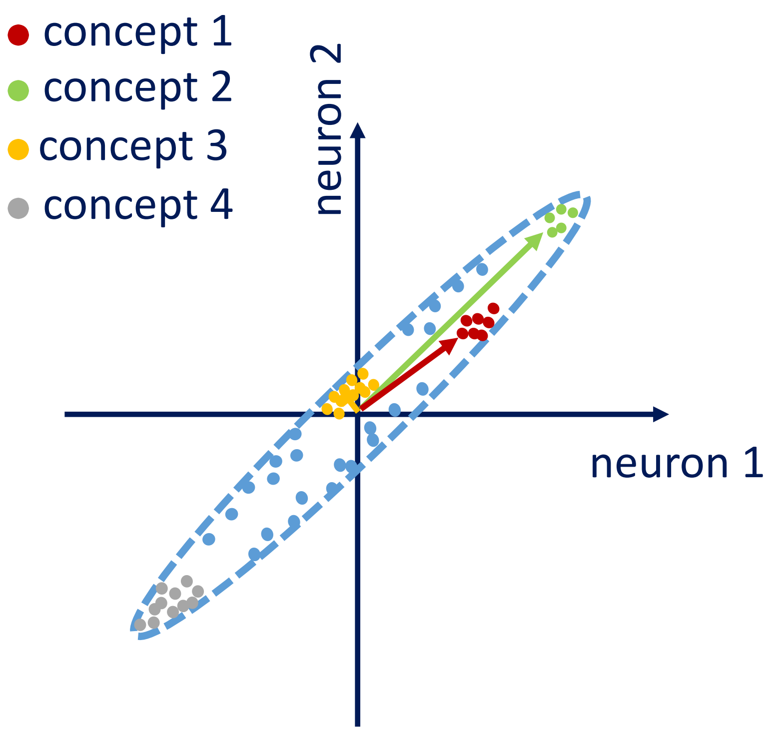


Duke

Concept Whitening

- Step 1: Whitening transformation
  - Decorrelate the latent space
  - Separate the concepts
- Step 2: Rotation transformation
  - Align the concepts to corresponding axes
  - Maintain the decorrelation property

Concept Whitening

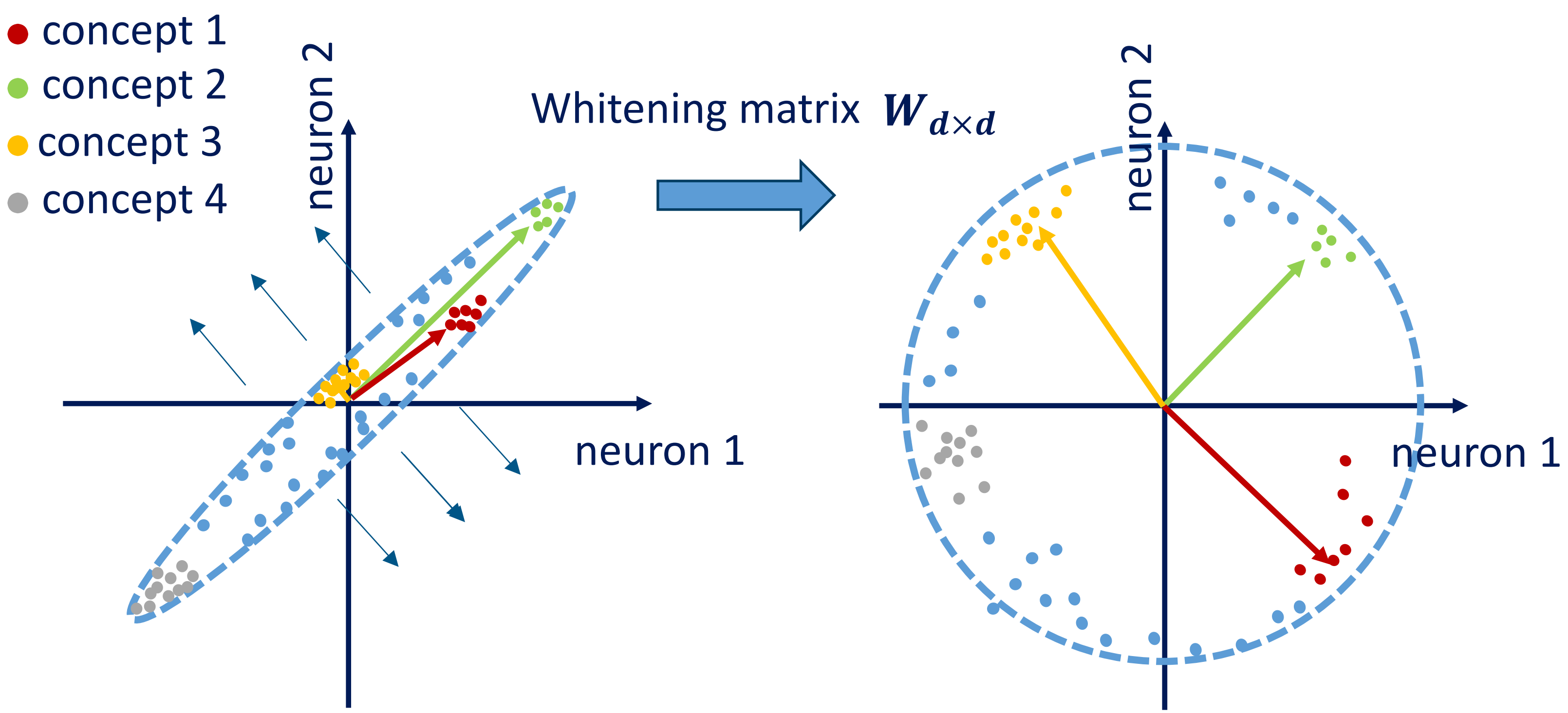


Mean-centered latent features

$$\widetilde{\boldsymbol{Z}}_{d\times n} = \boldsymbol{Z}_{d\times n} - \boldsymbol{\mu}\,\boldsymbol{1}^{\boldsymbol{T}}_{n\times 1}$$

Duke

Concept Whitening



Mean-centered latent features

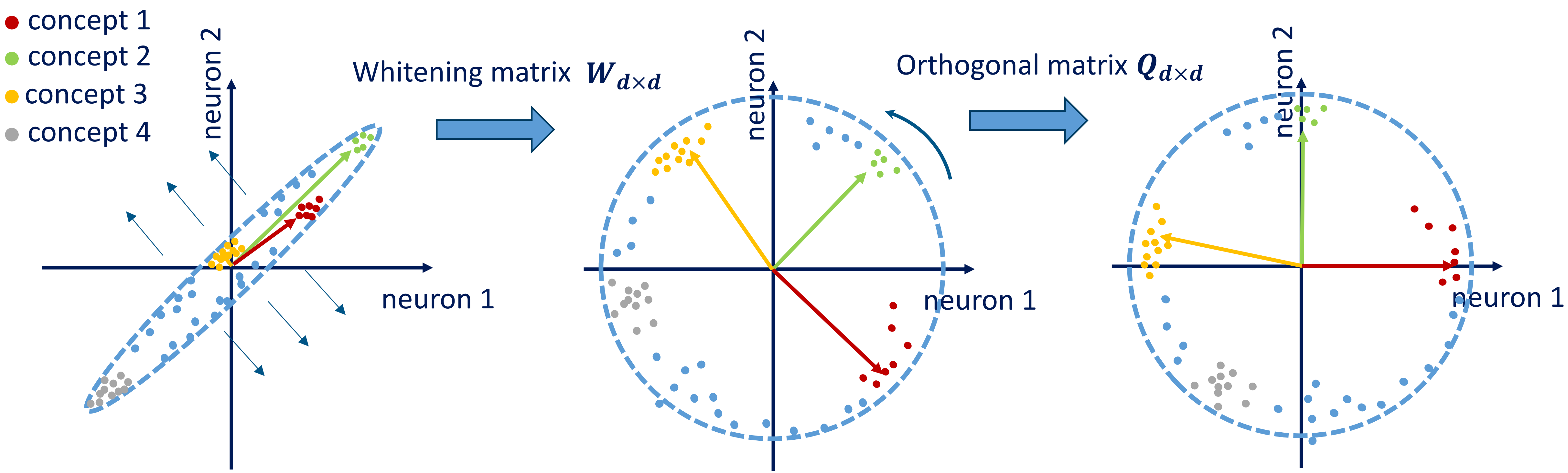
$$\tilde{\boldsymbol{Z}}_{d\times n} = \boldsymbol{Z}_{d\times n} - \boldsymbol{\mu}\, \boldsymbol{1}^{T}_{n\times 1}$$

After whitening transformation

$$\boldsymbol{W}\,\tilde{\boldsymbol{Z}}$$

$\boldsymbol{W}$ should obey $\boldsymbol{W}^{T}\boldsymbol{W} = \left(\frac{\tilde{\mathbf{Z}}\,\tilde{\mathbf{Z}}^{\mathrm{T}}}{n}\right)^{-1}$

Concept Whitening



Mean-centered latent features

$$\tilde{Z}_{d\times n} = Z_{d\times n} - \mu\, \mathbf{1}_{n\times 1}^T$$

After whitening transformation

$$W\tilde{Z}$$

W should obey $W^T W = \left(\frac{\tilde{Z}\tilde{Z}^{\mathrm{T}}}{n}\right)^{-1}$

After rotation transformation

$$Q^T W \tilde{Z}$$

$W' = Q^T W$ is also a whitening matrix

Learning the parameters

- Sample mean $\boldsymbol{\mu}$ and whitening matrix $\boldsymbol{W}$
  - Training phase: compute on the fly, support back-propagation (Huang et al)
  - Testing phase: exponential moving average of mini-batches (Ioffe & Szegedy)
- Orthogonal matrix $\boldsymbol{Q}$
  - maximizing concept activation under orthogonality constraint

$$\max_{\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_k} \sum_{j=1}^{k} \frac{1}{n_j} \boldsymbol{q}_j^T \boldsymbol{W} \boldsymbol{Z}_{c_j} \mathbf{1}_{n_j \times 1}$$
$$s.t. \boldsymbol{Q}^T \boldsymbol{Q} = \boldsymbol{I}_d$$

$\boldsymbol{Z}_{c_j}$: samples of concept j

$\boldsymbol{W} \boldsymbol{Z}_{c_j}$: after whitening

$\boldsymbol{q}_j^T \boldsymbol{W} \boldsymbol{Z}_{c_j}$: projection on axis j

$\frac{1}{n_j} \boldsymbol{q}_j^T \boldsymbol{W} \boldsymbol{Z}_{c_j} \mathbf{1}_{n_j \times 1}$: average activation

$\boldsymbol{Q}$ can be trained by gradient descent on Stiefel manifold (Wen & Yin, 2013)

Duke

What’s the cost of interpretability?

Duke

Main task performance

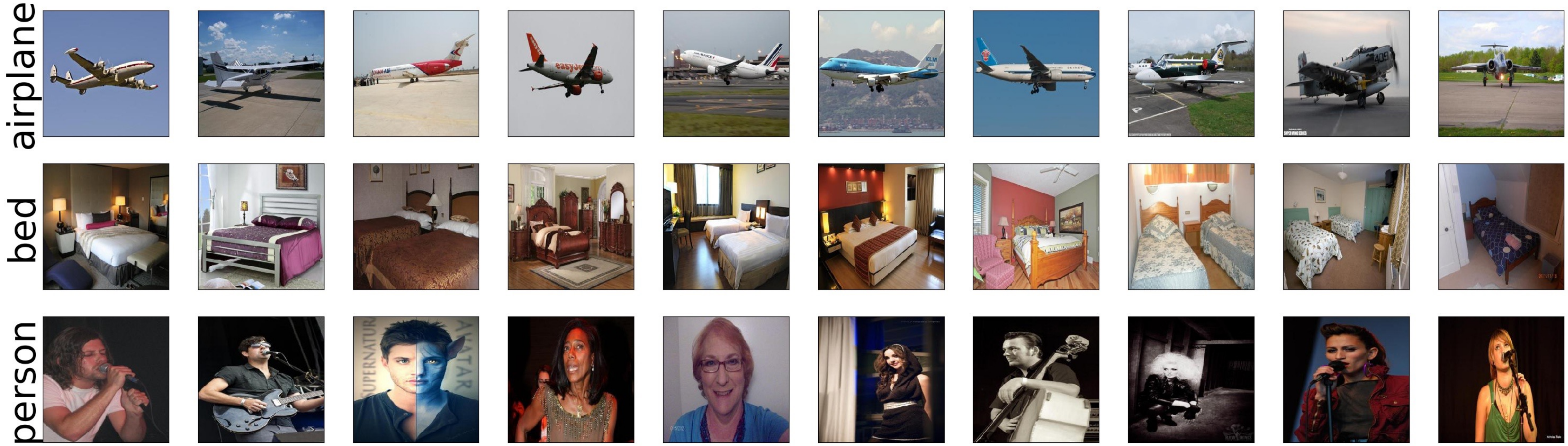
- accuracy is on par with standard CNNs
  - different datasets, backbone architectures, layers, #concepts
- warm-start from pretrained model
  - replace BN with CW
  - *only one* additional epoch of further training

What do the learned concepts look like?

Visualize the concept axes

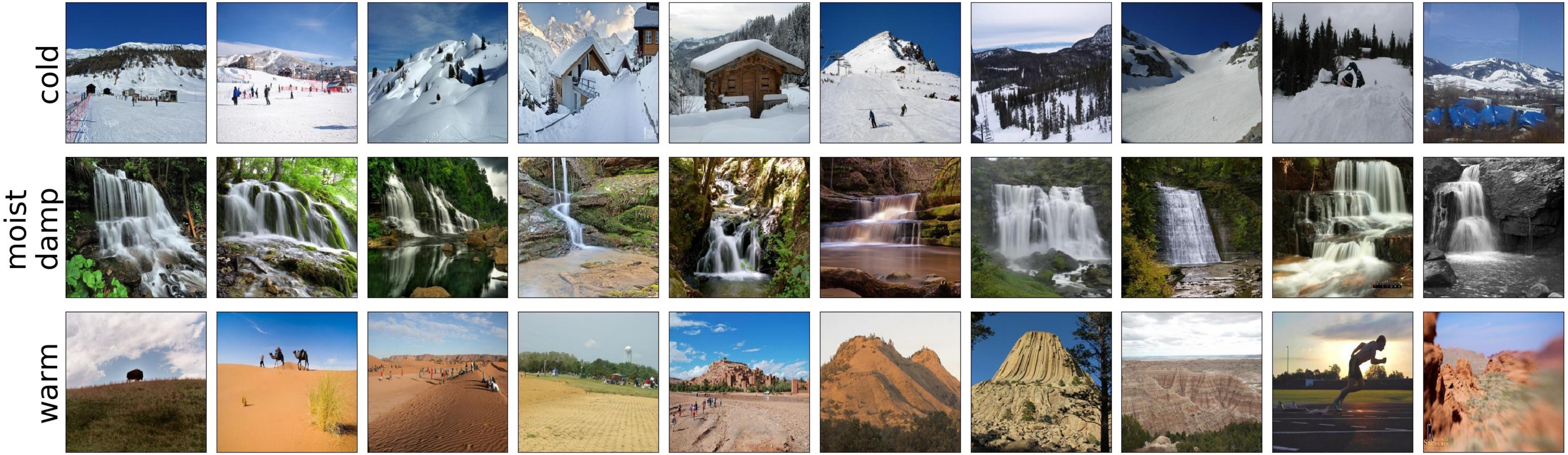
Most activated

16^{th} layer



Visualize the concept axes

- Not only objects - weather



Duke

Visualize the concept axes

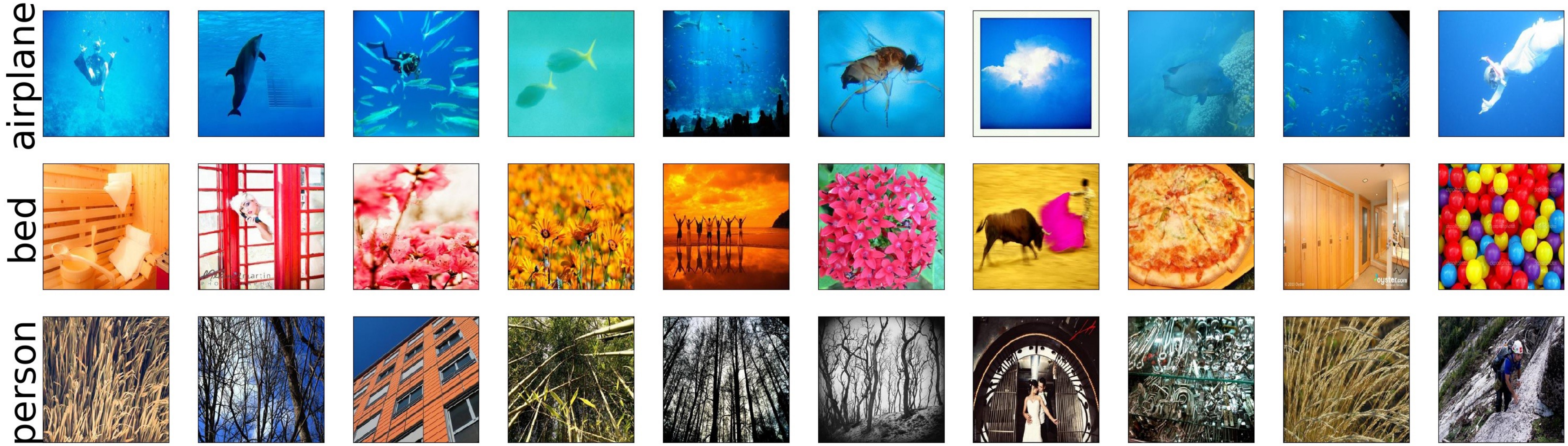
- Not only objects - material

Duke

Visualize the concept axes

Most activated

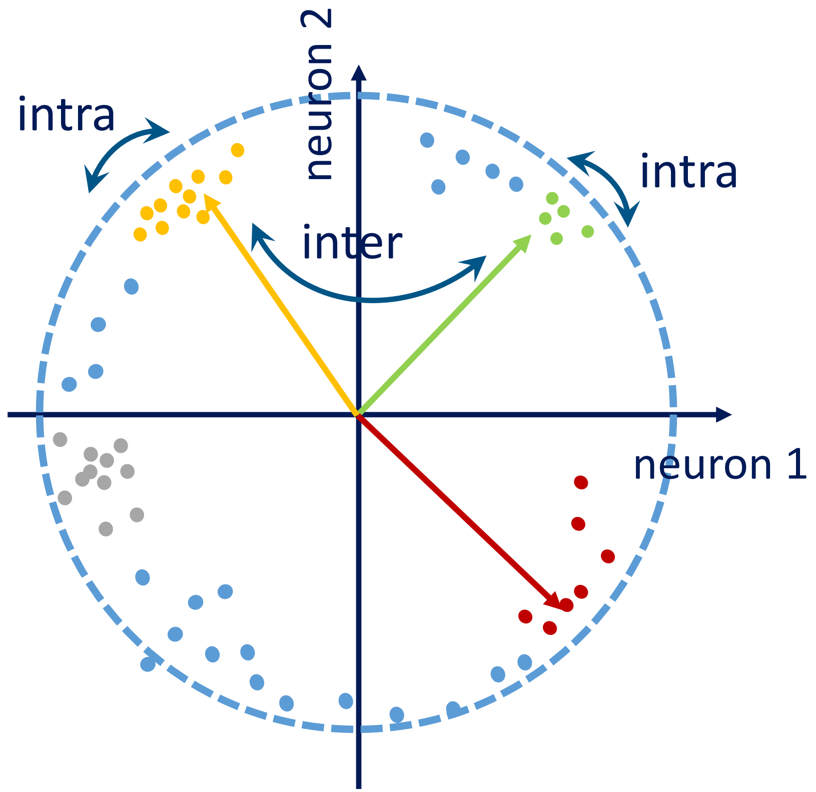
2^{nd} layer



How to quantitatively measure the quality of the learned concepts?

Concept separation

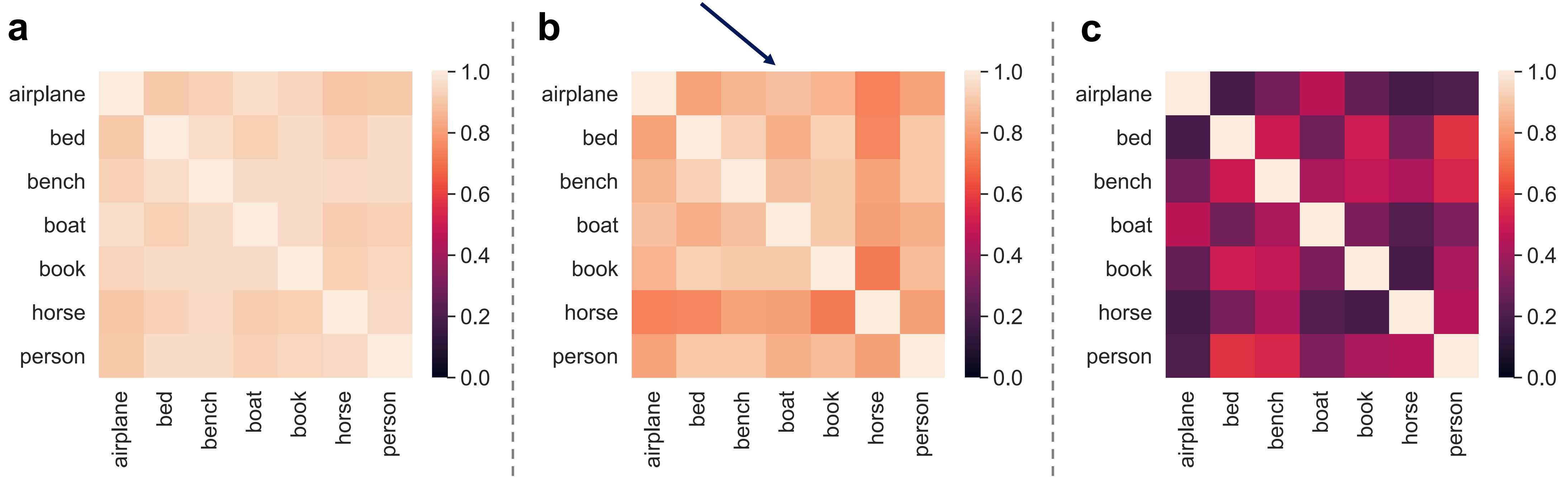
$$\text{inter- intra-concept ratio} = \frac{\text{avg cos similarity between concept } i \text{ and concept } j}{\sqrt{\text{avg cos similarity concept } i}\sqrt{\text{avg cos similarity concept } j}}$$



Duke

Concept separation

directly build a concept classifier in the latent space

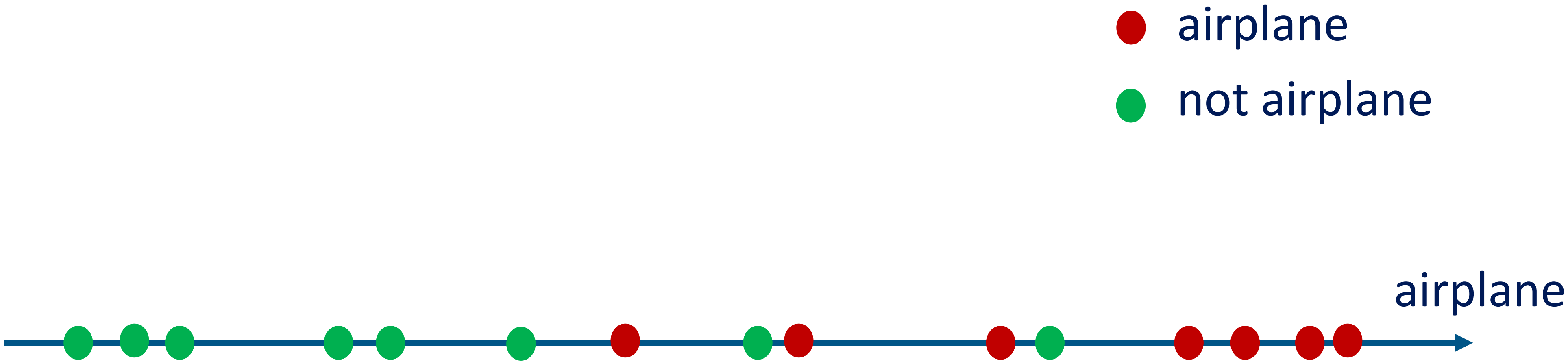


BatchNorm
(avg inter-sim = 0.74)

auxiliary concept classification loss
(avg inter-sim = 0.74)

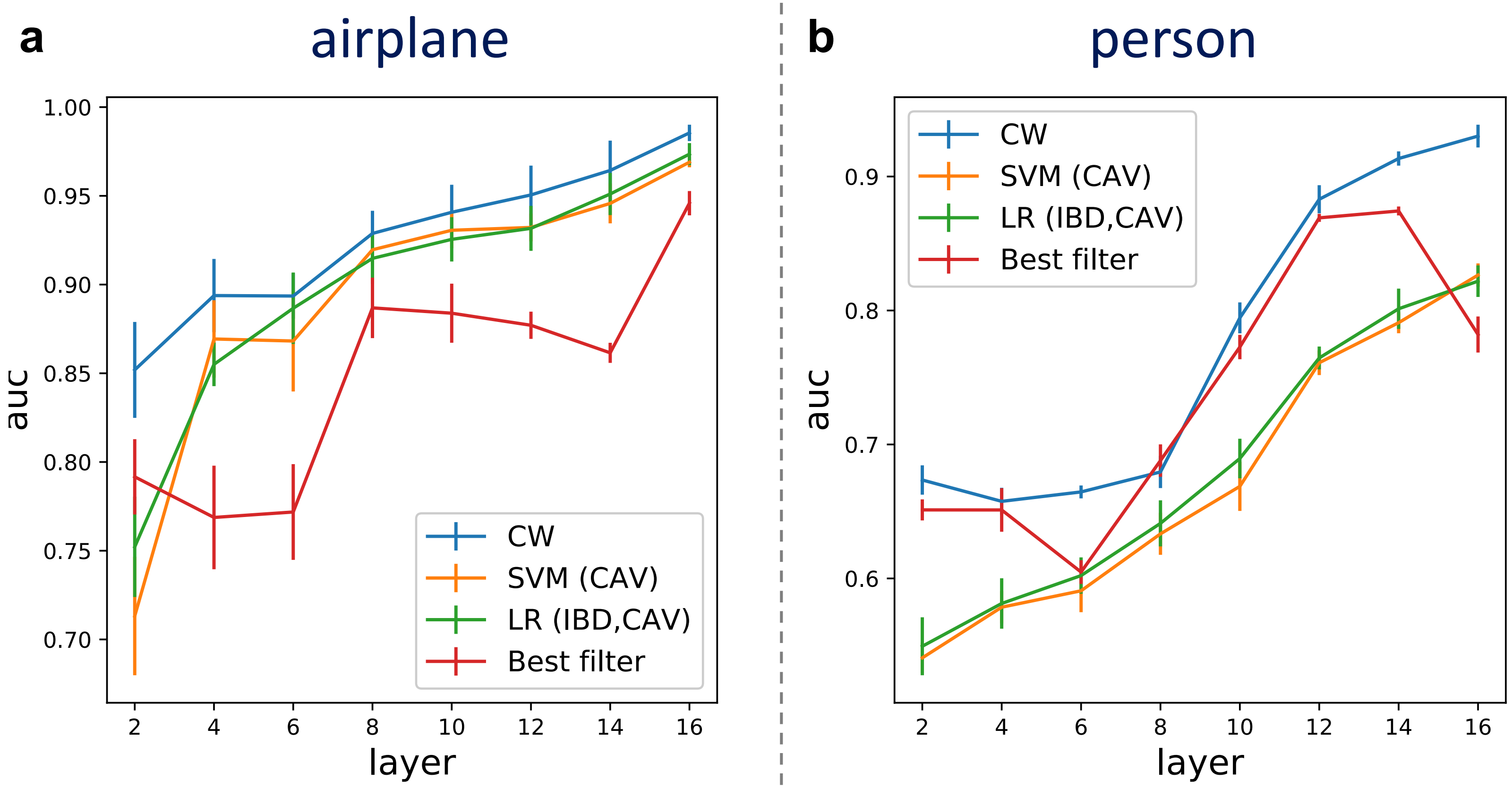
CW
(avg inter-sim = 0.05)

Concept purity



AUC of the activation measures concept purity

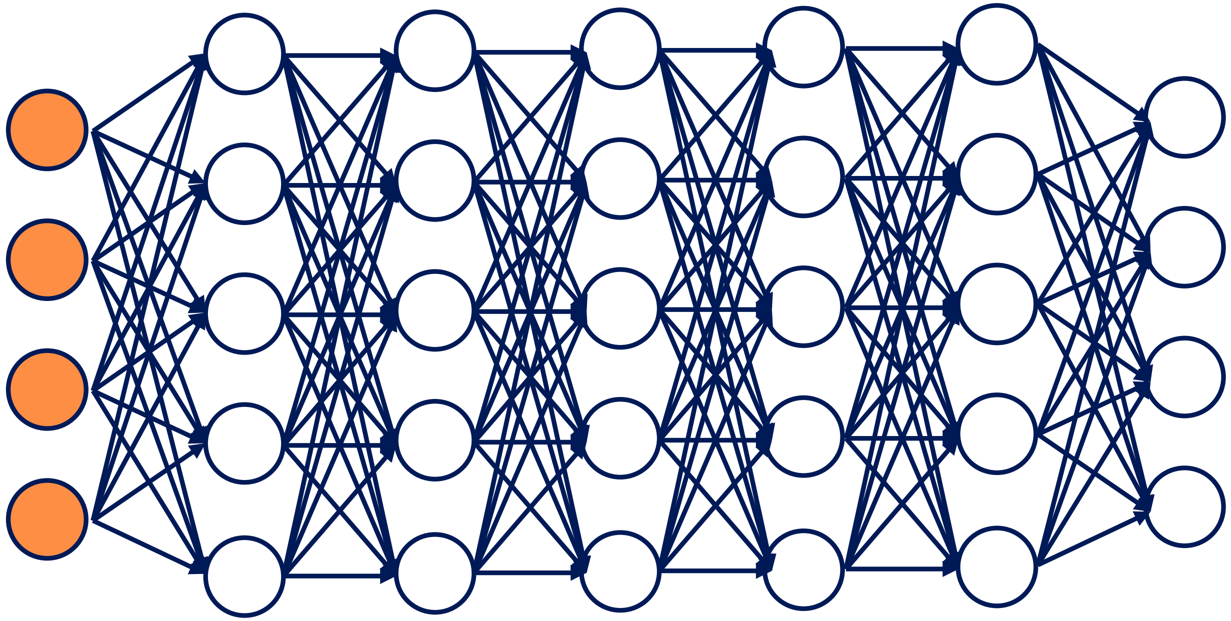
Concept purity

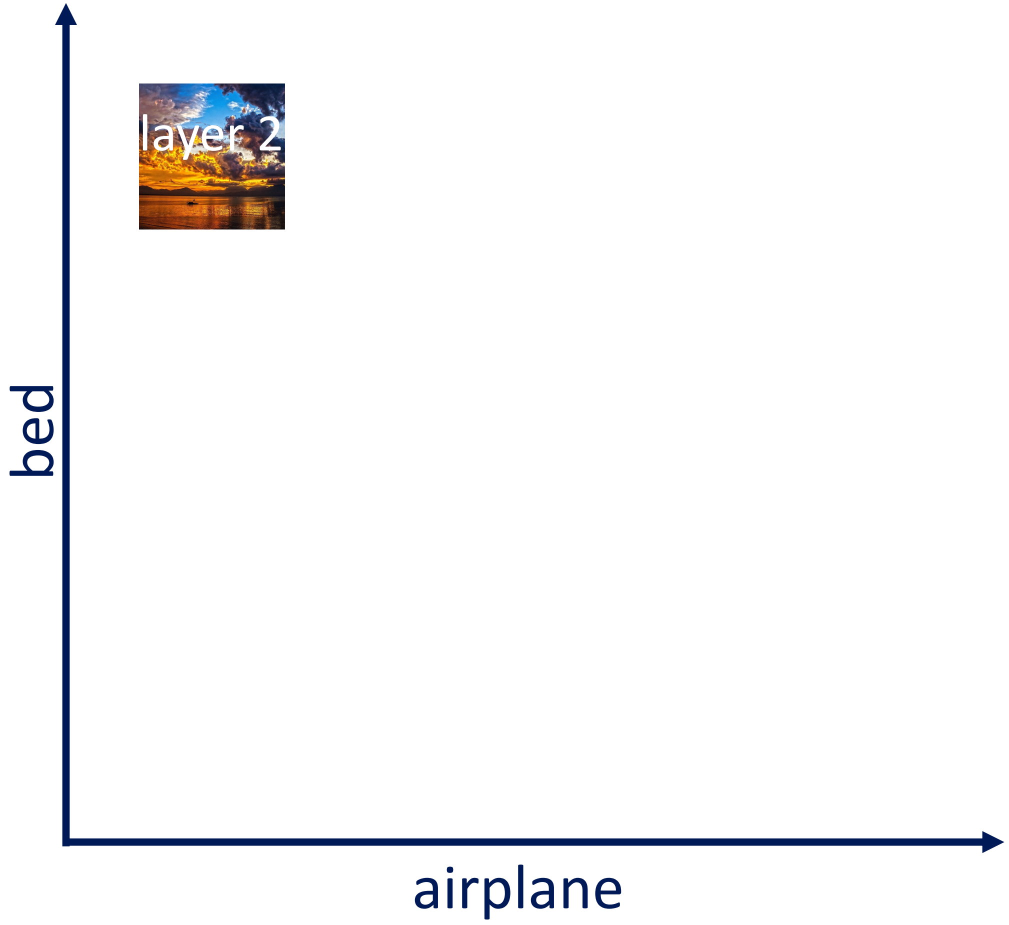


Duke

What can we use this model for?

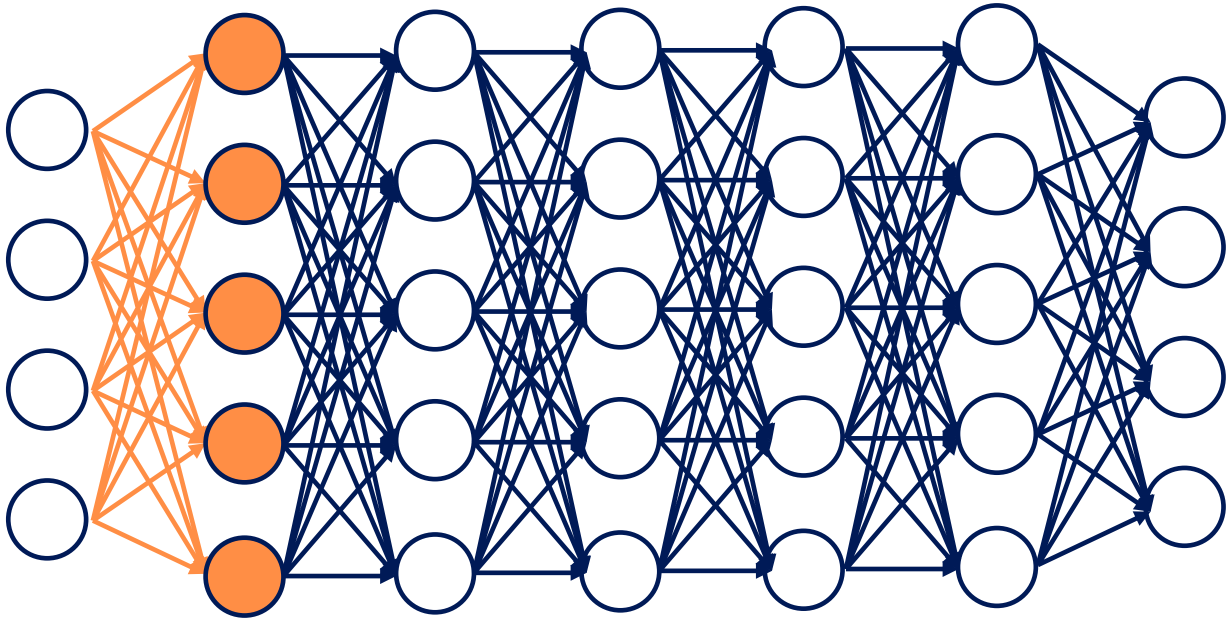
Reasoning process

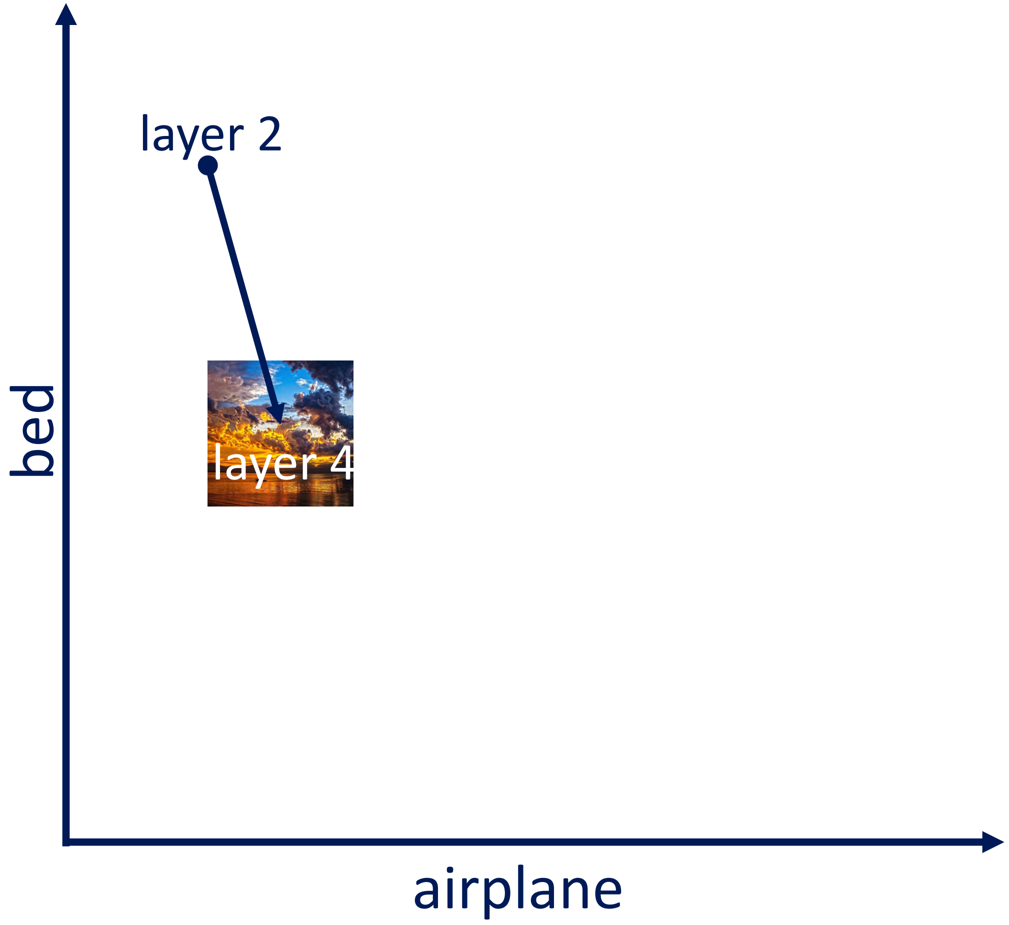




Duke

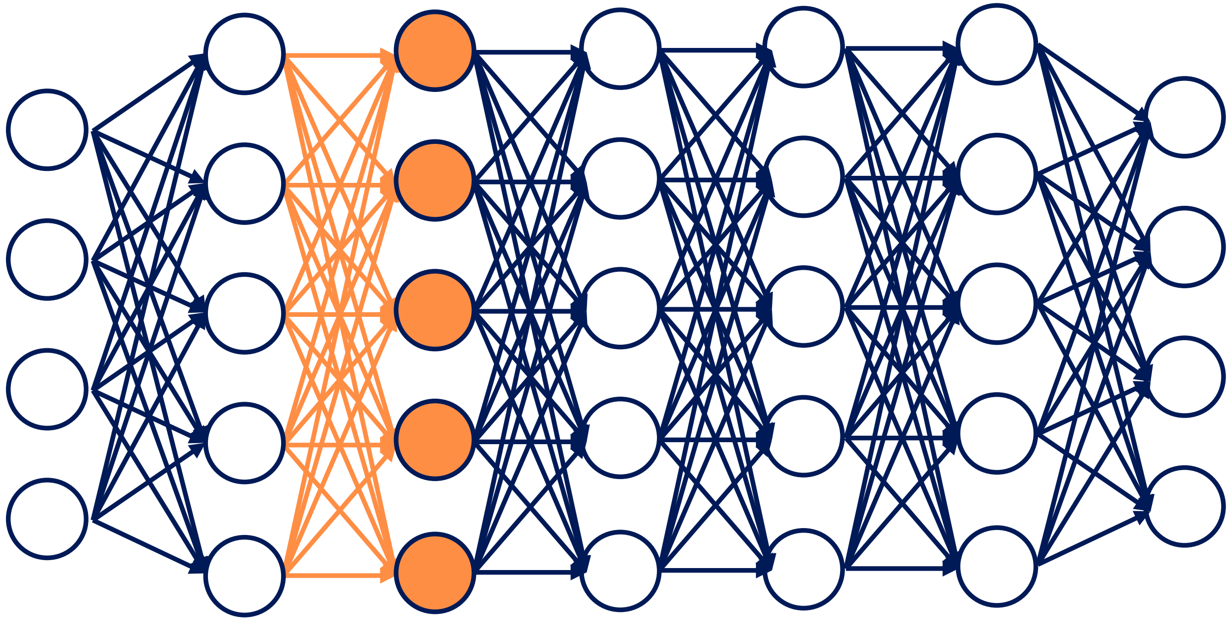
Reasoning process

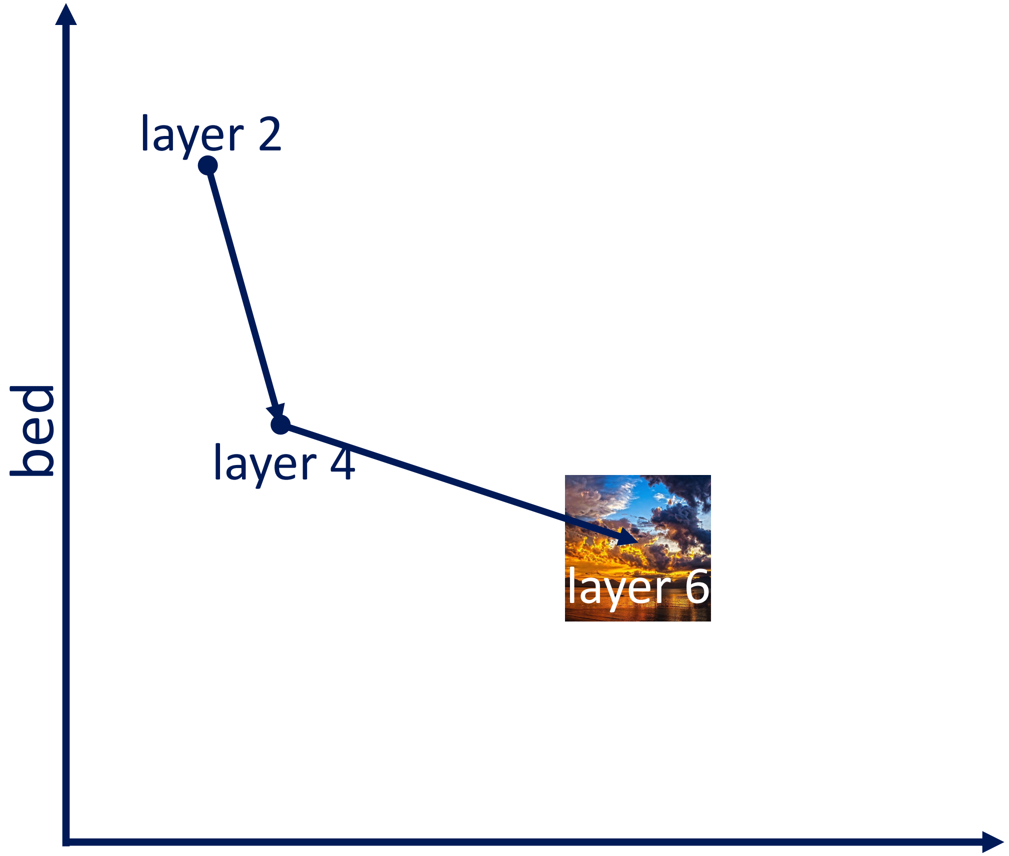




Duke

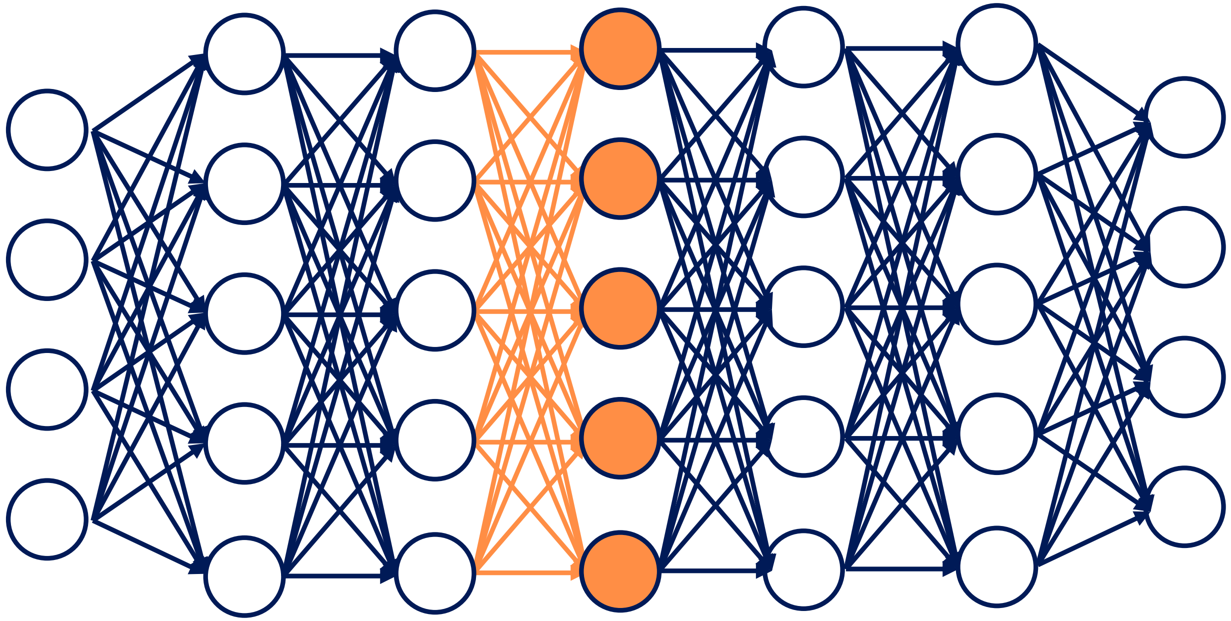
Reasoning process

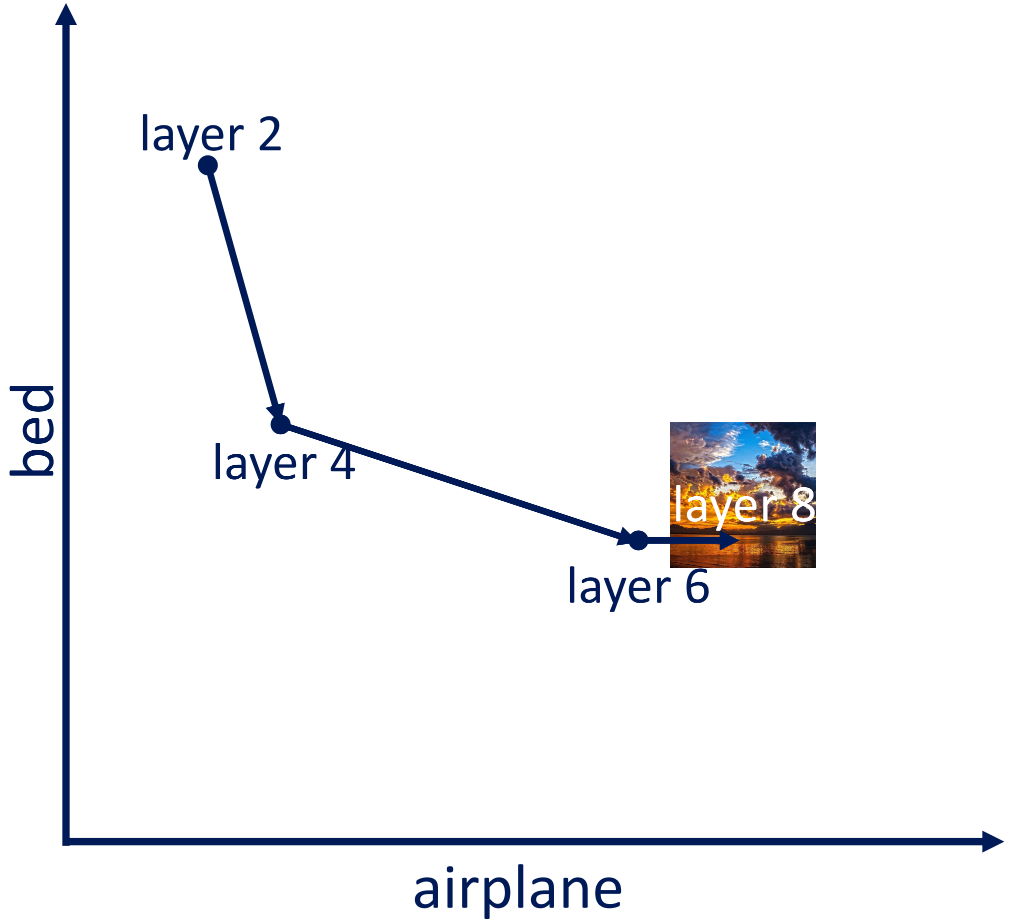




Duke

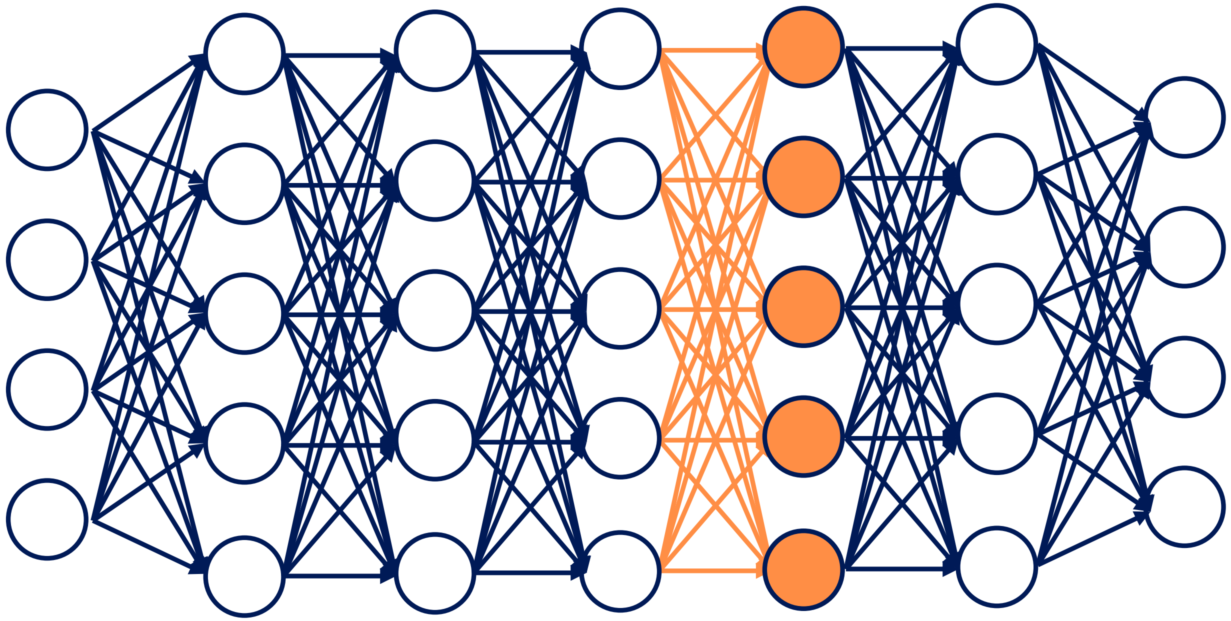
Reasoning process

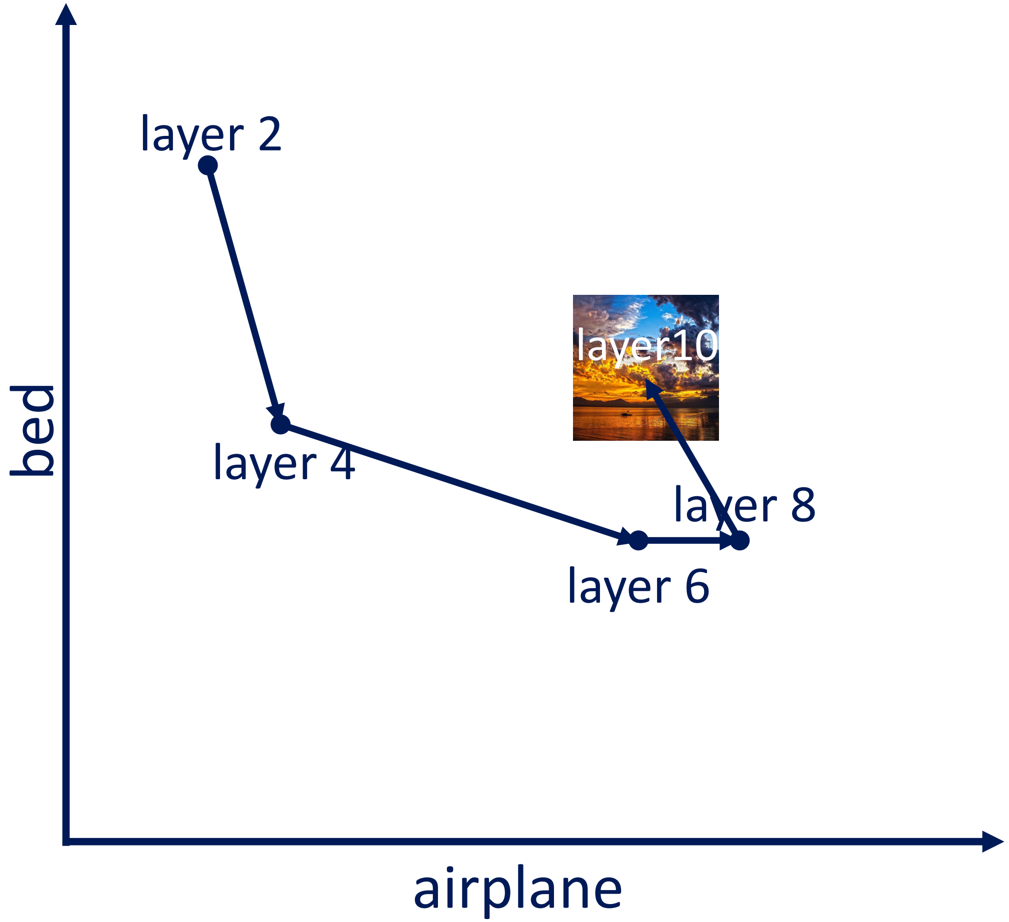




Duke

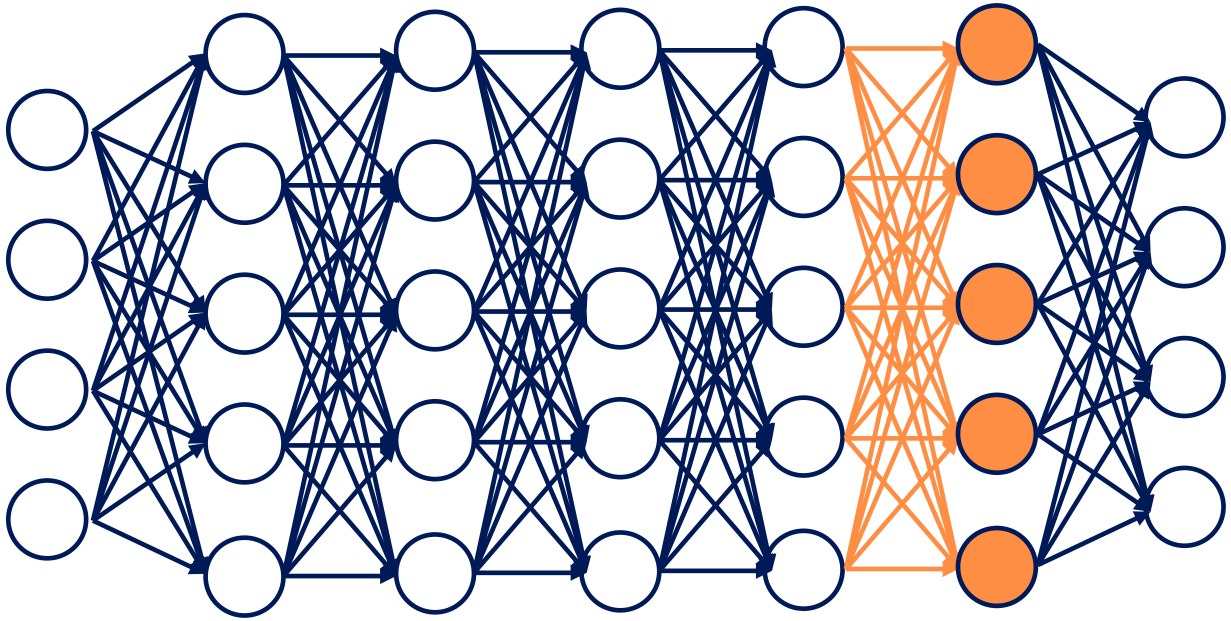
Reasoning process

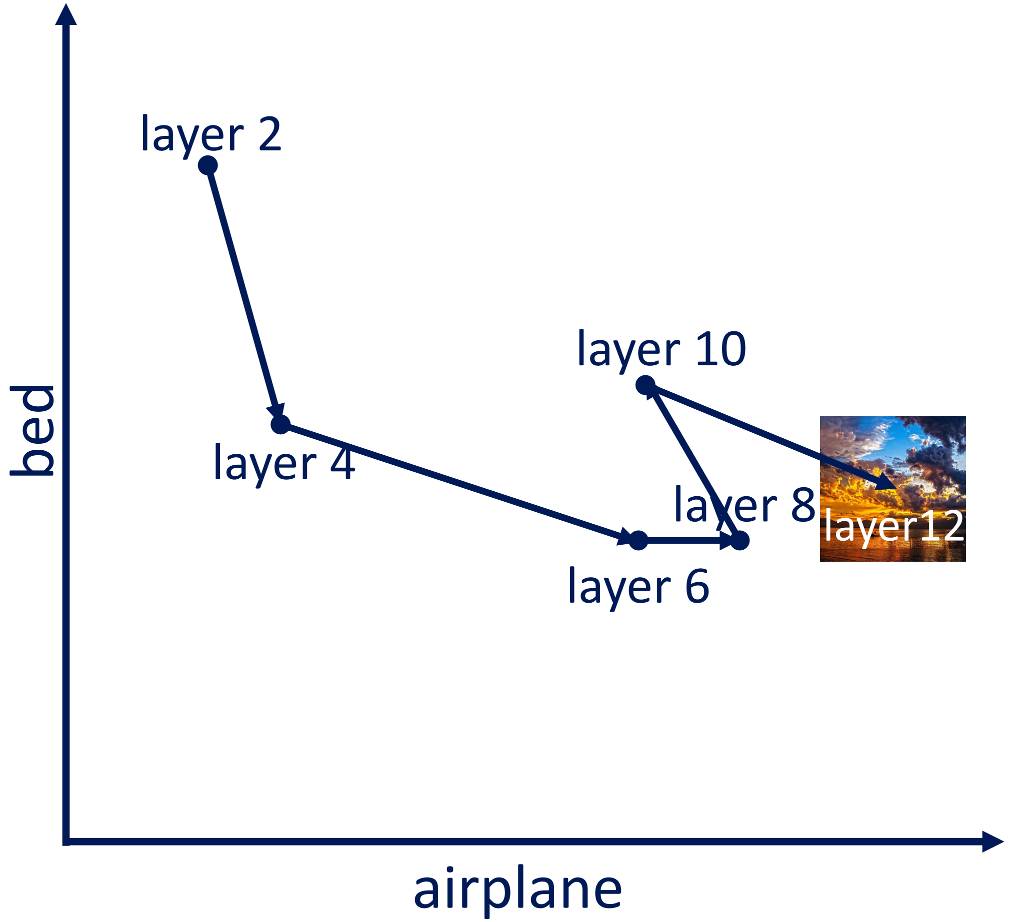




Duke

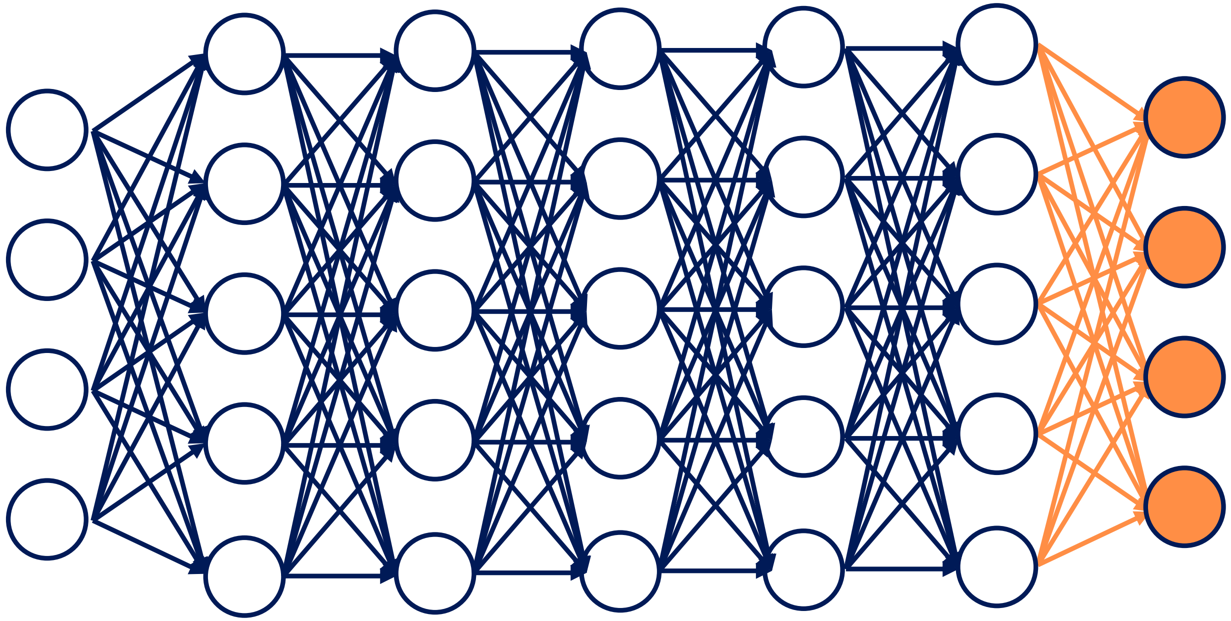
Reasoning process

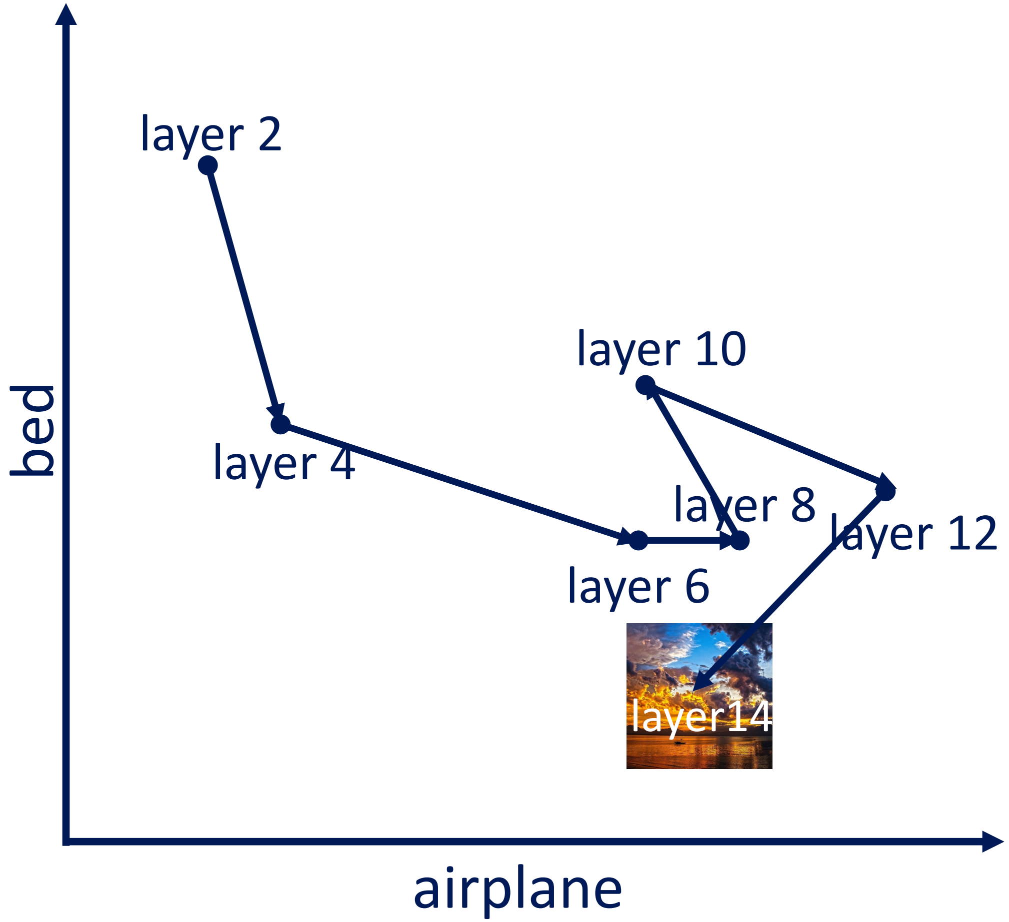




Duke

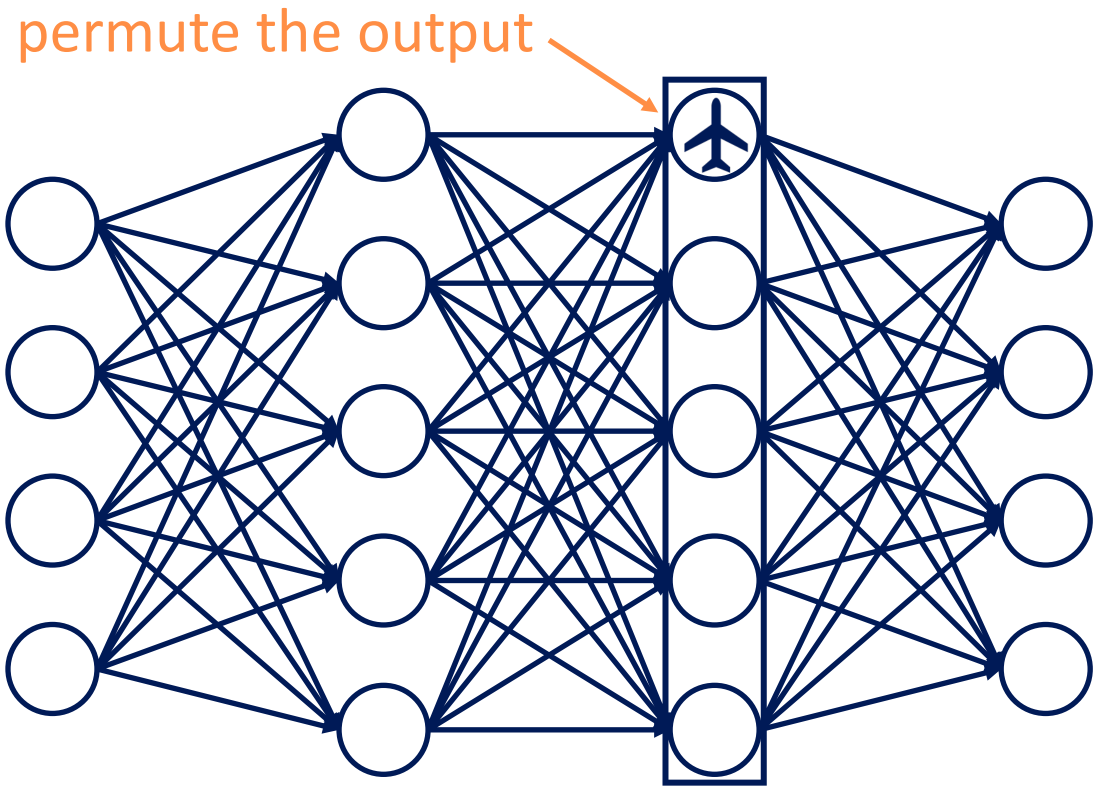
Reasoning process





Duke

Concept importance



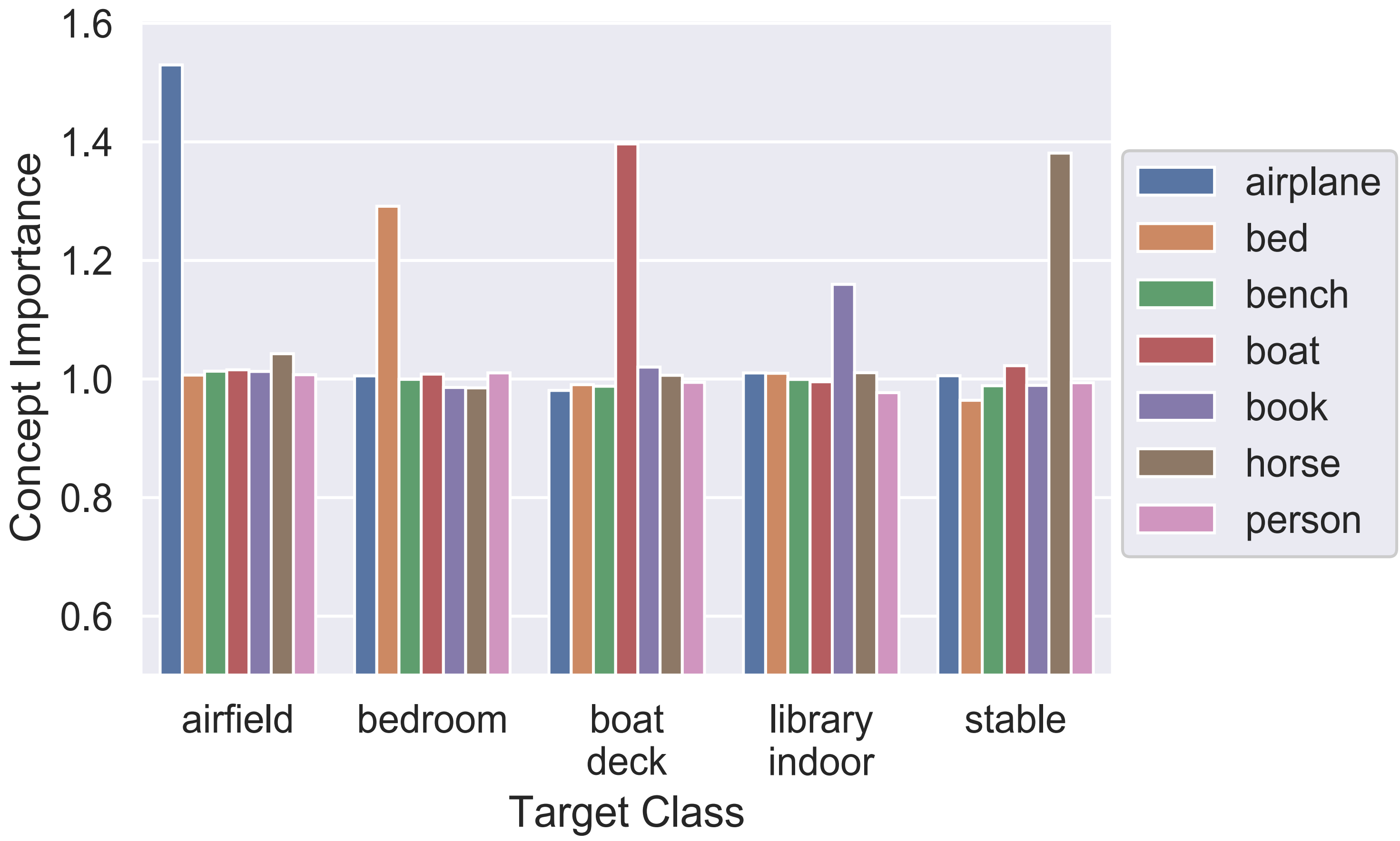
- Variable importance of axis j

$$CI_j = \frac{\mathrm{e}_{\mathrm{switch}}^{(j)}}{e_{\mathrm{original}}}$$

Duke

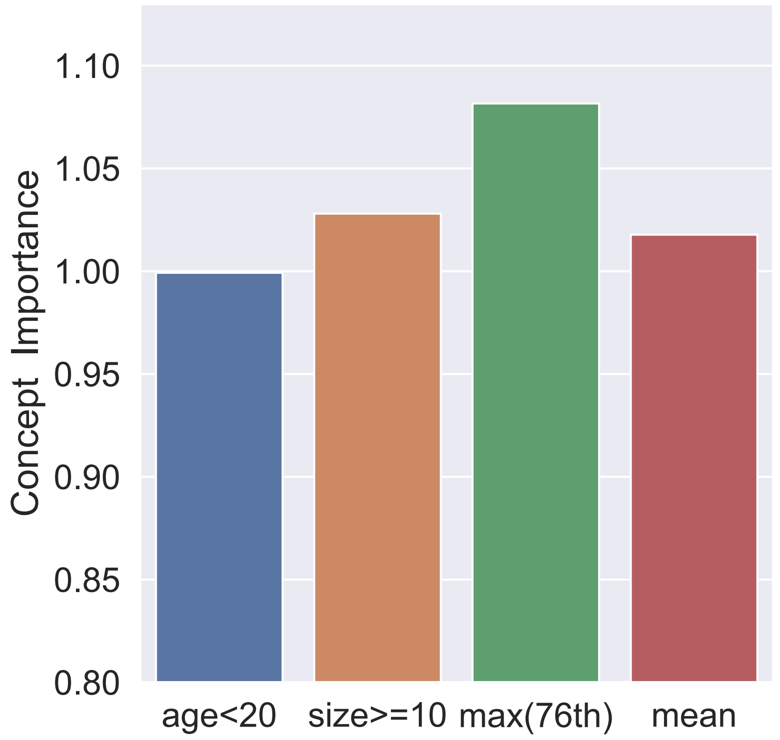
Concept importance

- Scene classification
  - Places365



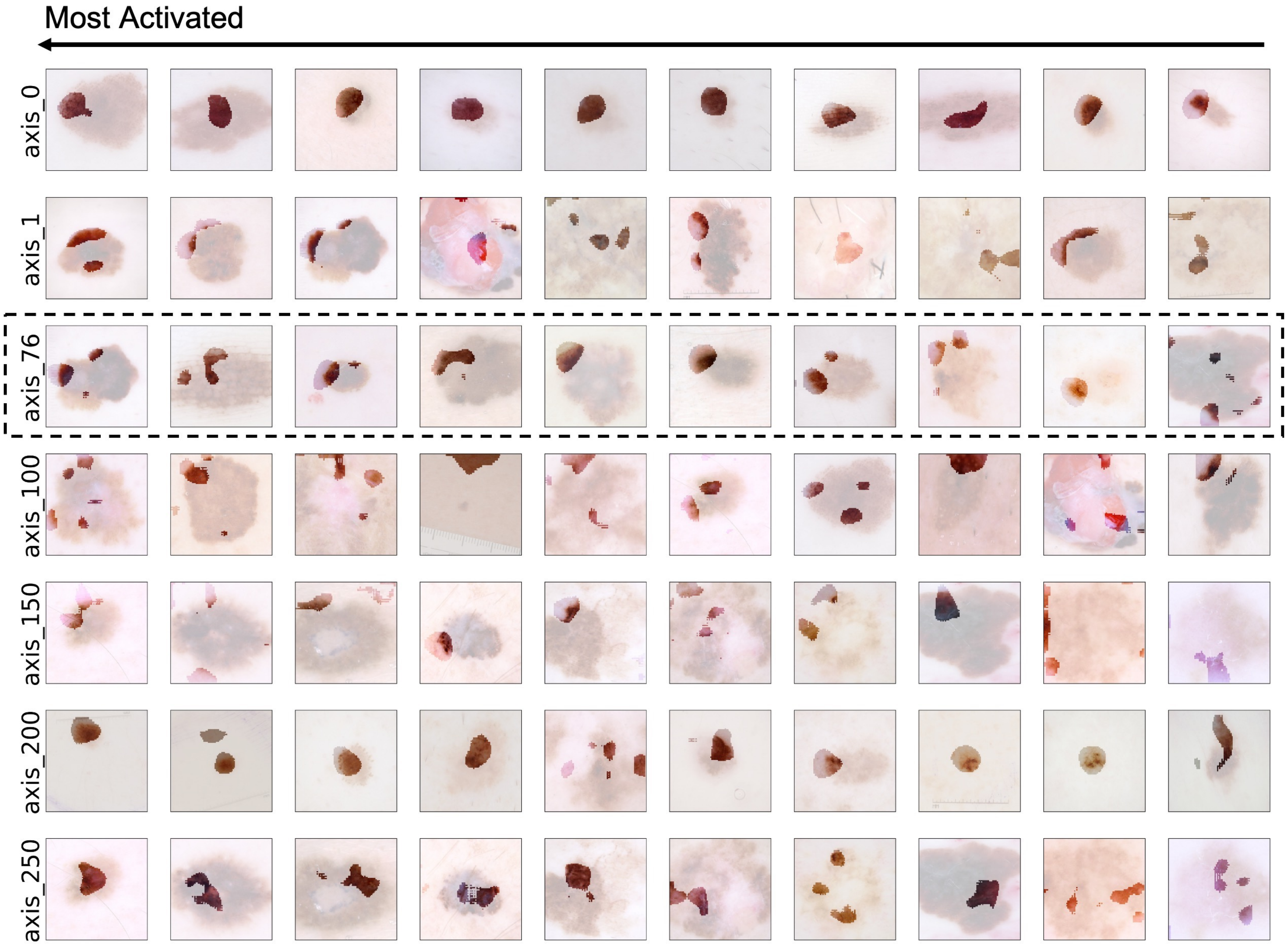
Concept importance

- Skin lesion malignancy
  - ISIC dataset
  - axis 1: age < 20
  - axis 2: size >= 10 mm
  - not most important



Duke

focused on boundaries

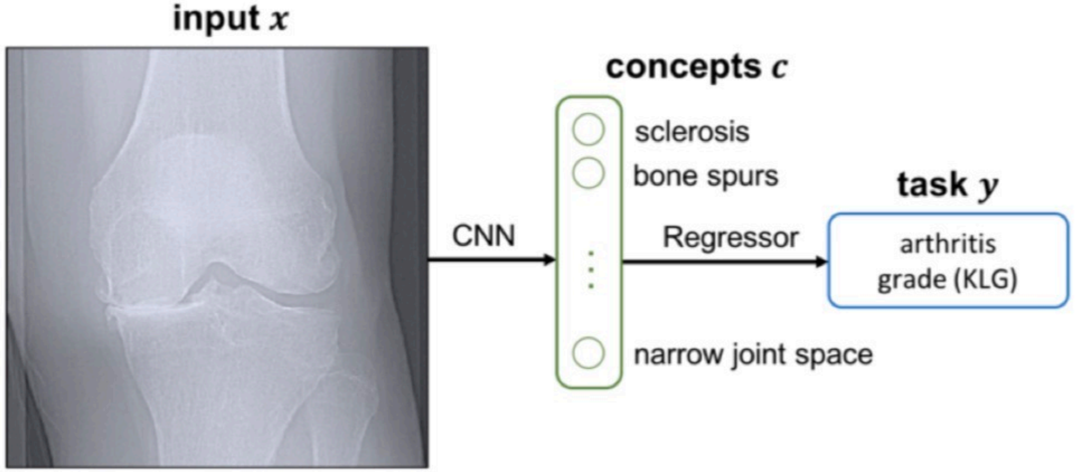


doctors also agree
boundary > size > age

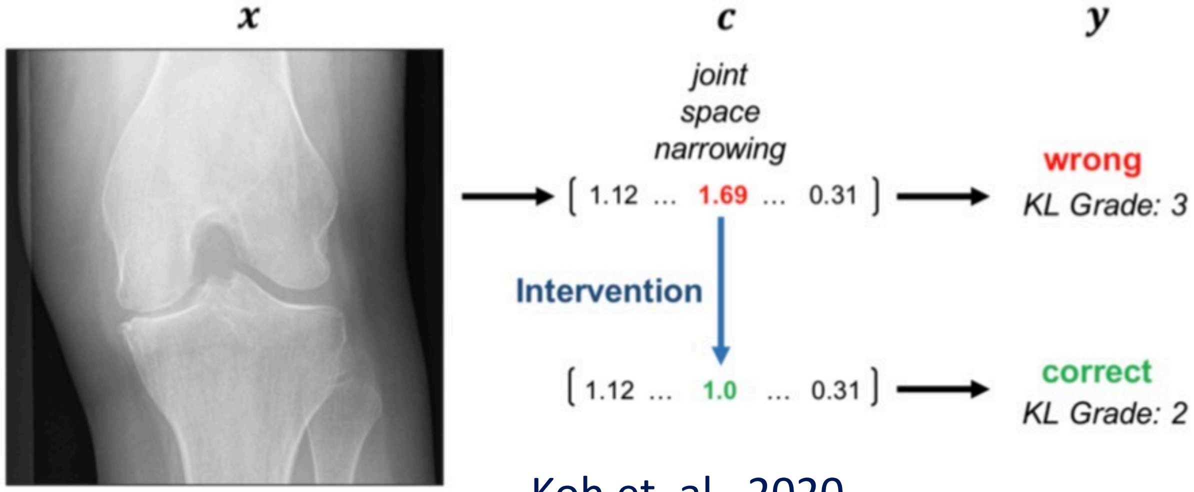
Duke

Model intervention and editing

- Concept Bottleneck Model (Koh et. al , 2020)
  - they didn't disentangle
  - concept-based models can do test-time intervention



doctors can change the model when it is wrong



Koh et. al , 2020

Duke

Summary: Concept Whitening

- ➢ Better interpretability

  - concepts are disentangled in the latent space

- ➢ No sacrifice in accuracy

  - accuracy is on par with standard CNNs

- ➢ Easy to use

  - warm-start from pretrained model requires only one additional epoch of further training

Duke

Links

- Nature Machine Intelligence paper

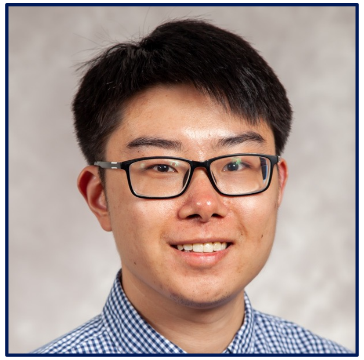
  - https://rdcu.be/cbOKj

- Code

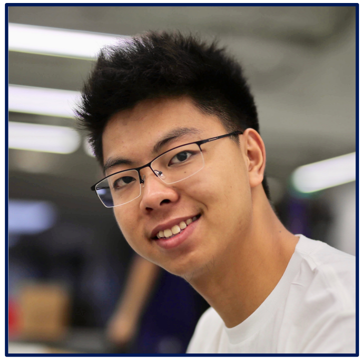
  - https://github.com/zhiCHEN96/ConceptWhitening

Duke

Thank you



Zhi Chen



Yijie Bei

Cynthia Rudin

Thank you to Carlo Tomasi, Guillermo Sapiro, Wilson Zhang, Lesia Semenova, Harsh Parikh, Chudi Zhong, and Oscar Li.

Funded by MIT-Lincoln Laboratory and the National Science Foundation

Duke