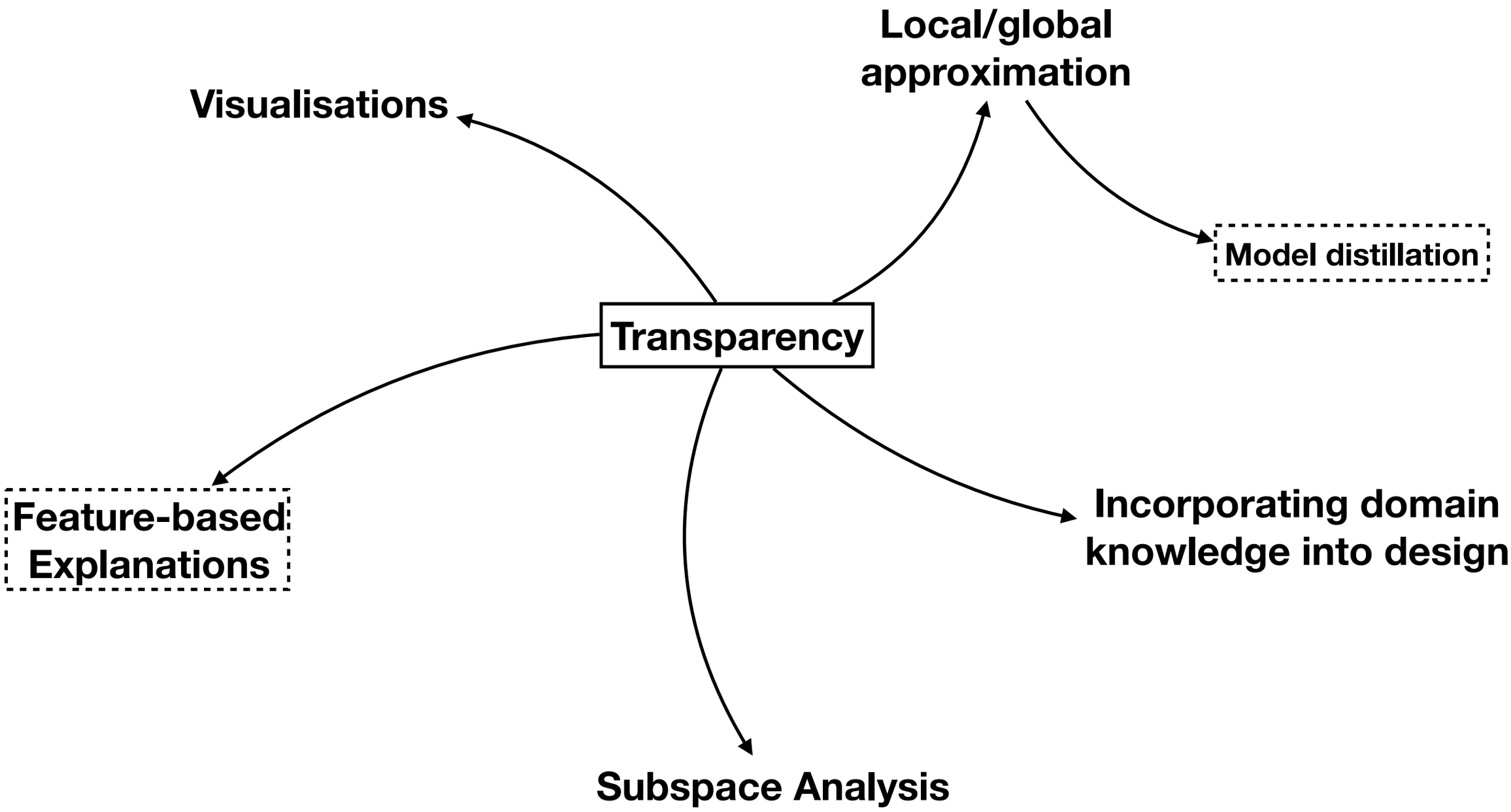
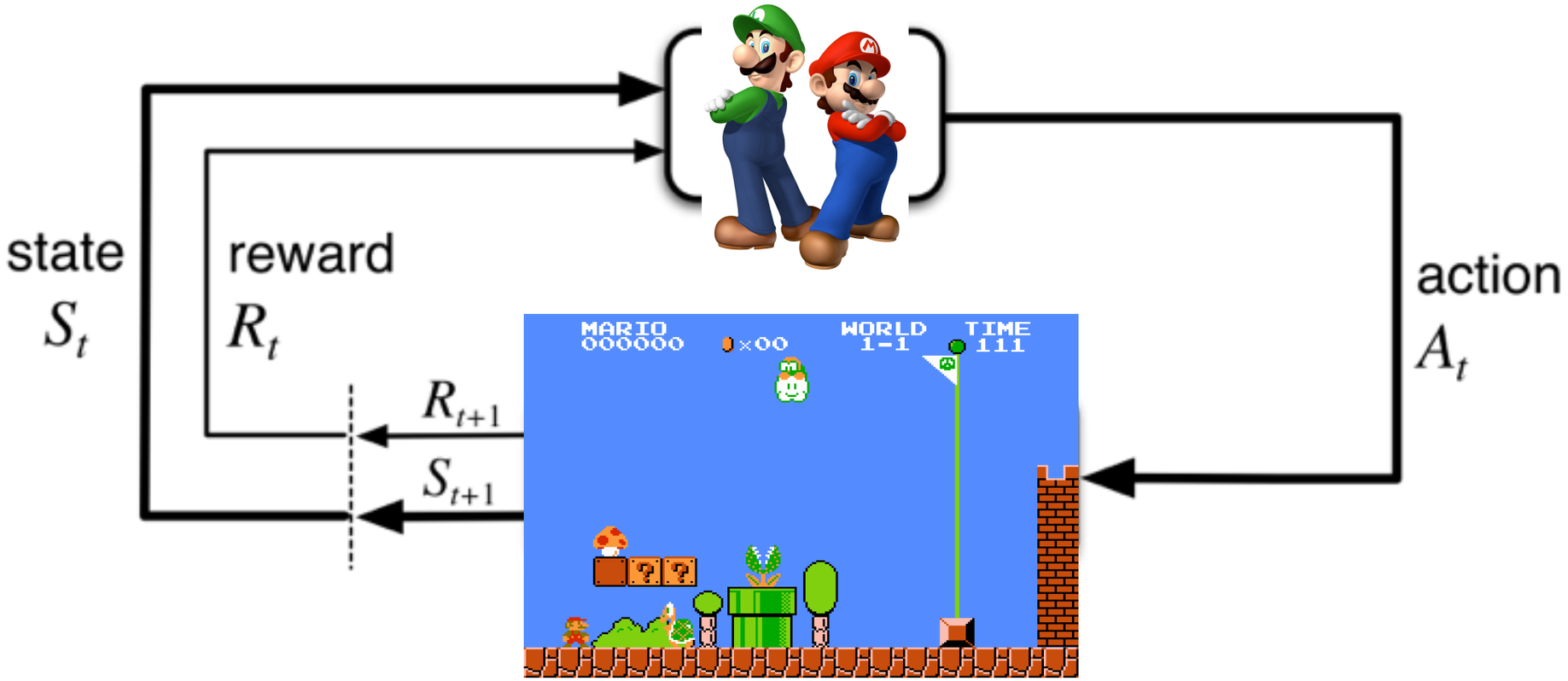
Generating Post-hoc Explanations for Neural Networks Through Concept Mapping

Siddharth Ravi, Vlado Menkovski, Henry Brighton

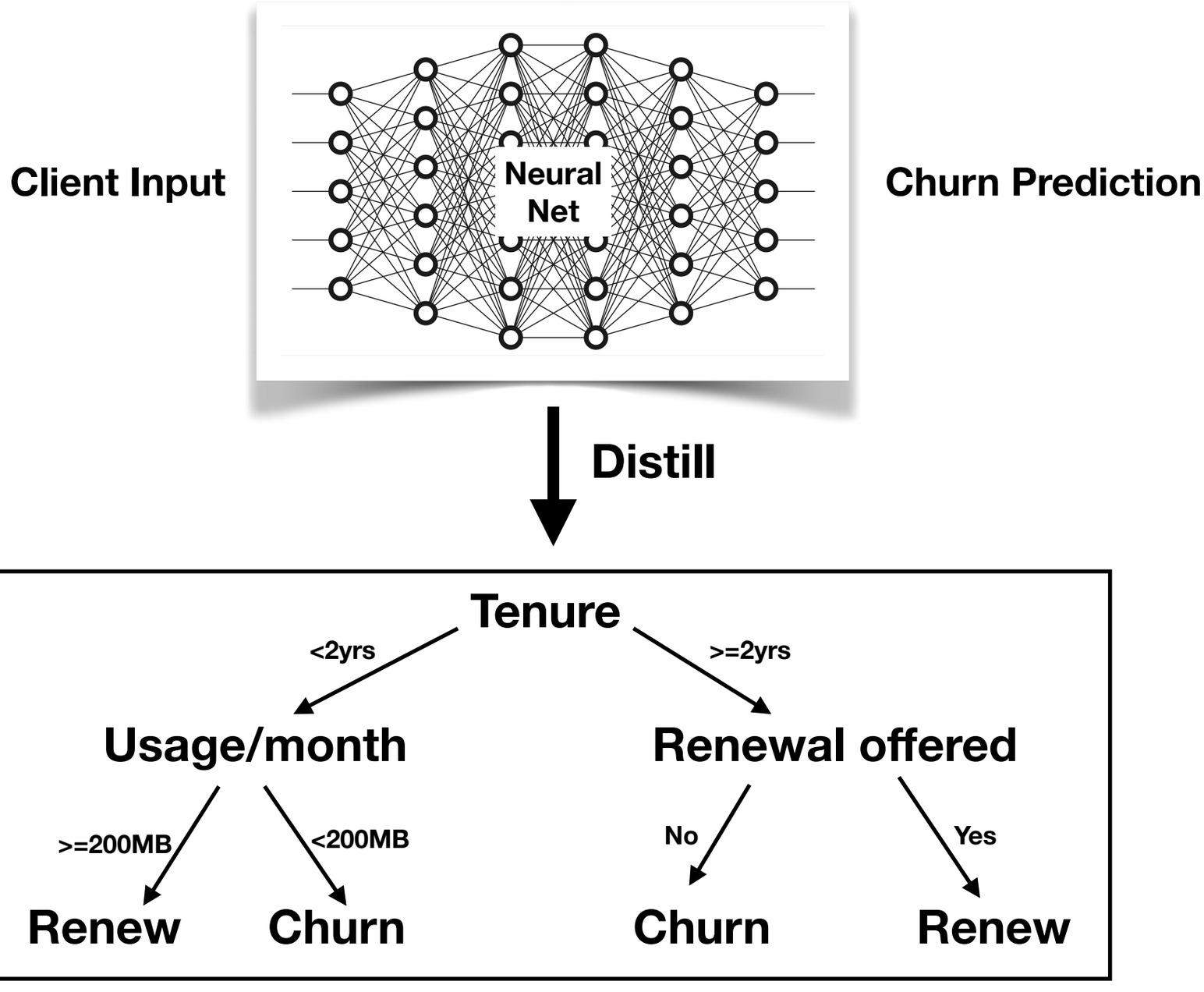
Ways to improve transparency



Reinforcement Learning

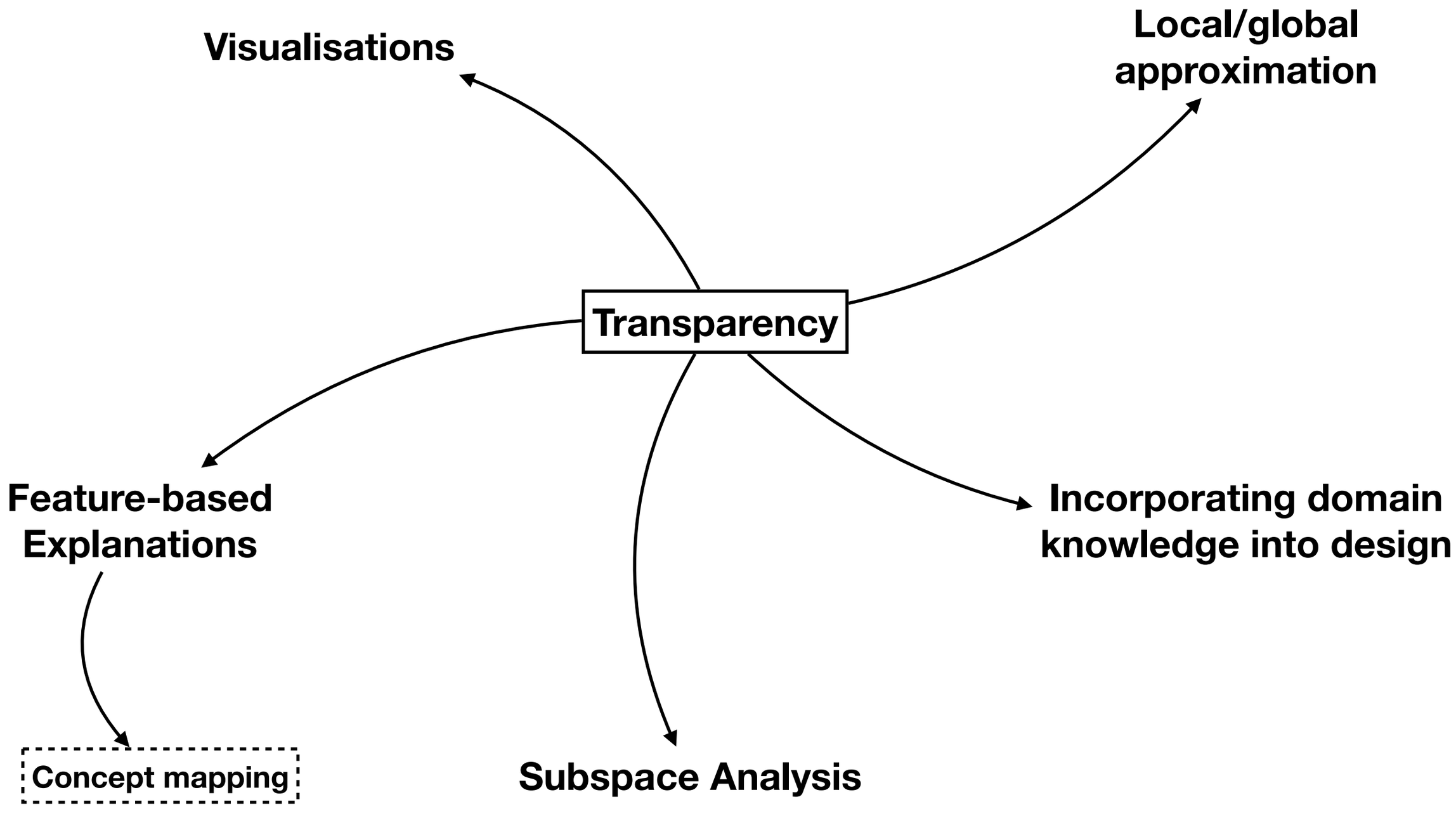


Model Distillation

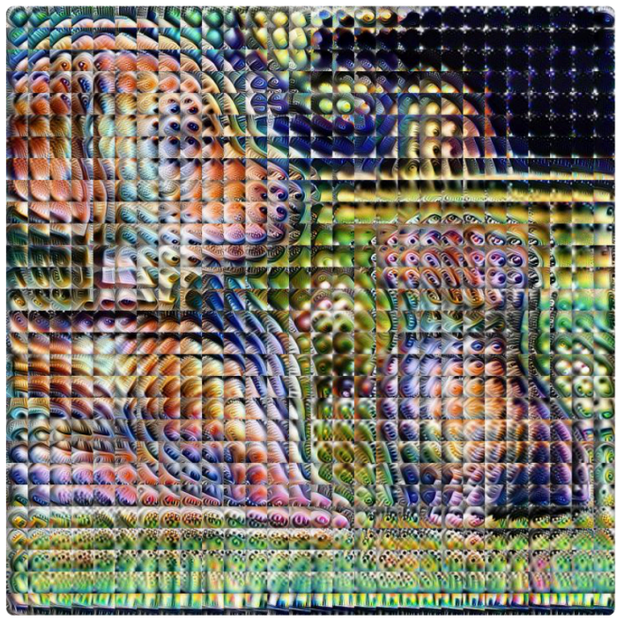
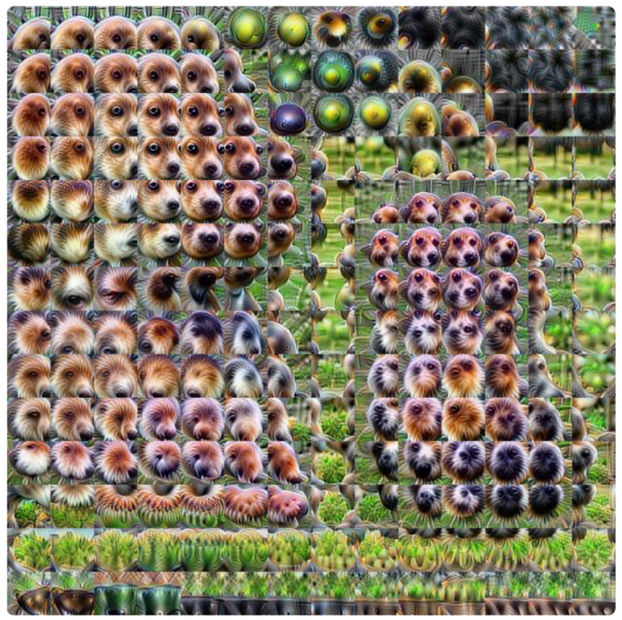
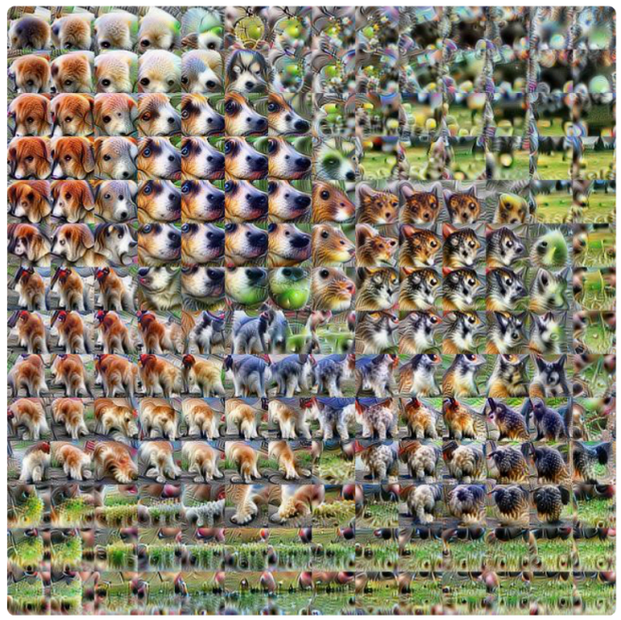
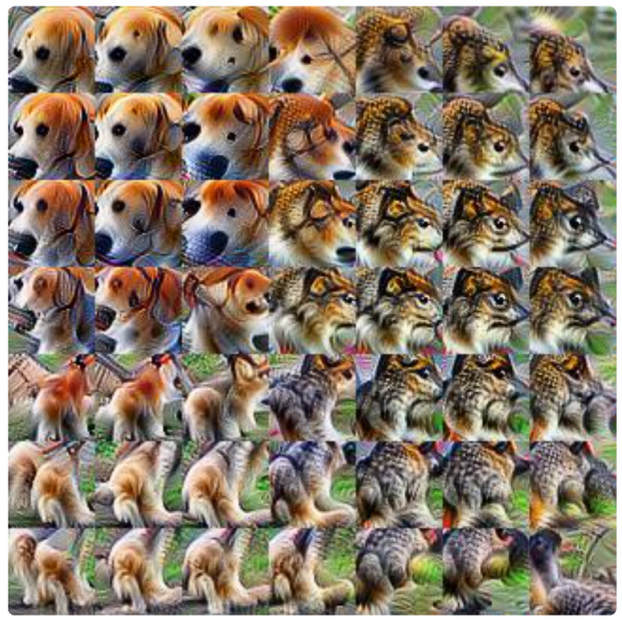


Reference: *Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." arXiv preprint arXiv:1503.02531 (2015).*

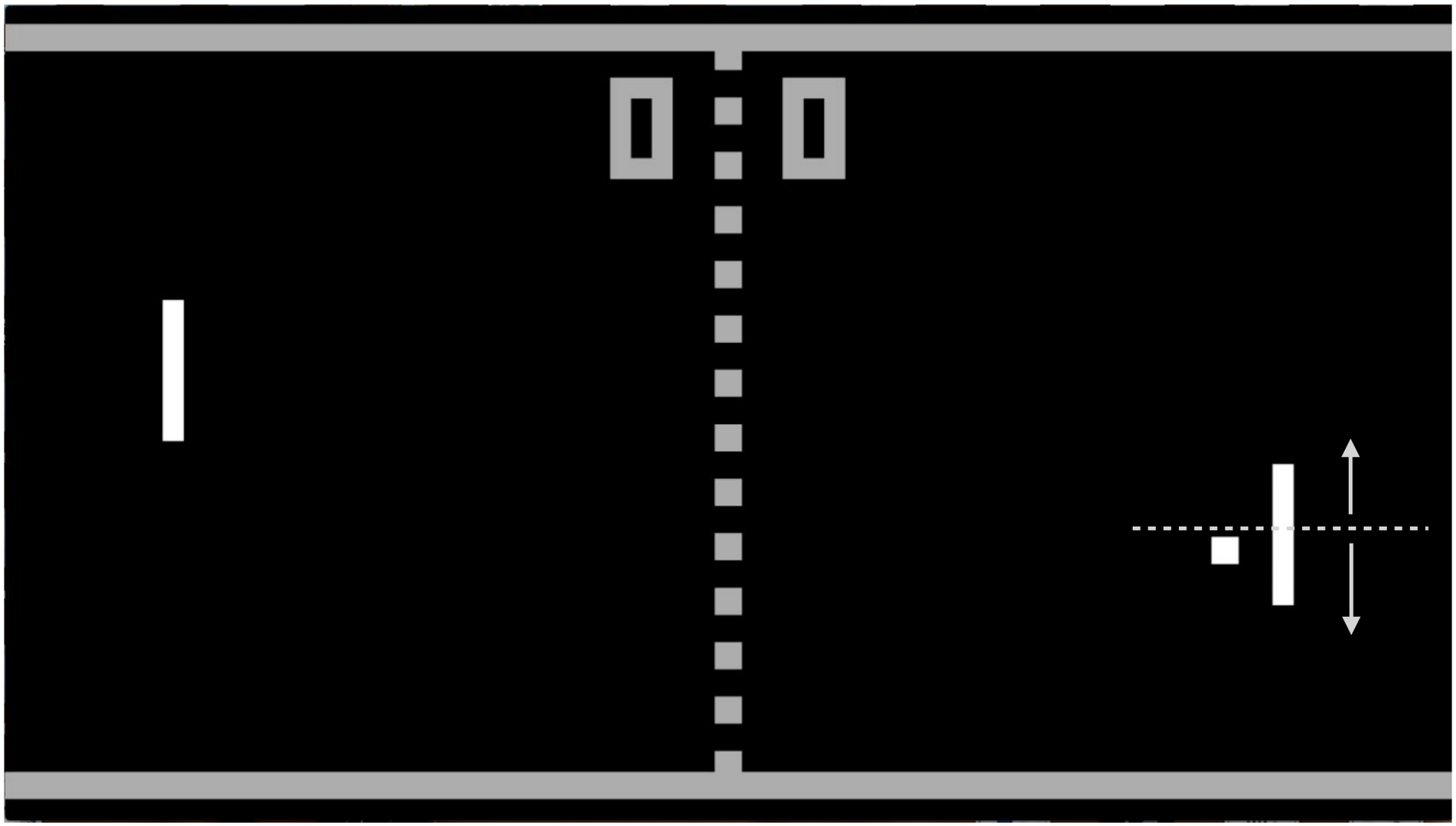
Many ways to improve transparency



Neural Activity Maps

Pong

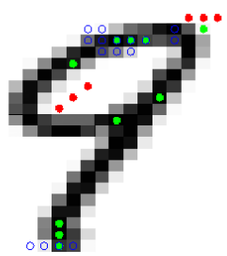


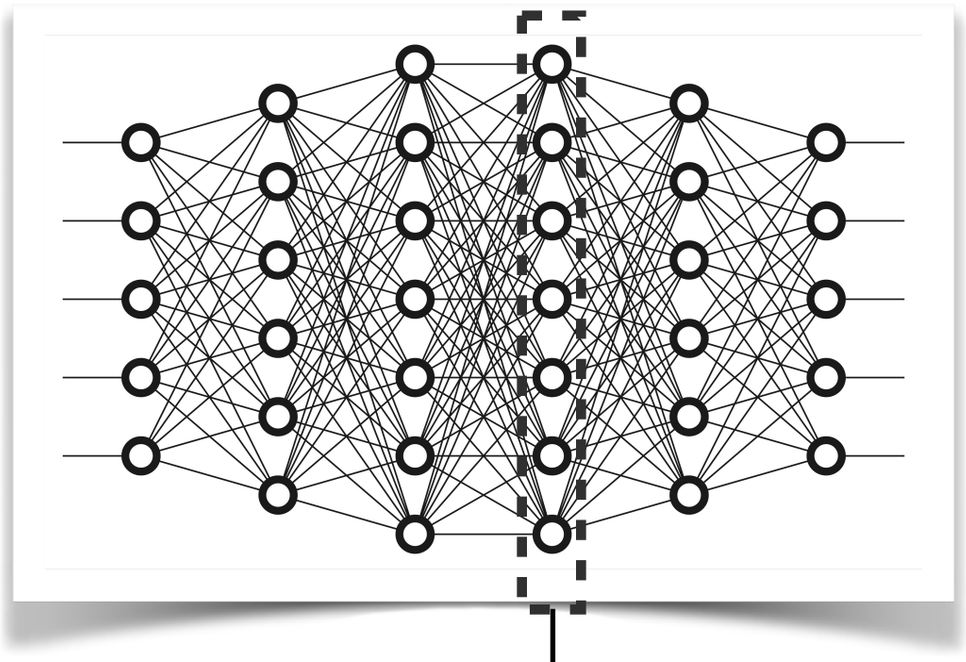
Concept Space

Target Output	Ball below bisector	Ball above bisector
Move down	1	0
Move up	0	1

Problem statement: Can we create linear maps from latent state representations in neural nets to a concept space?

Concept mapping for MNIST





[0-9]

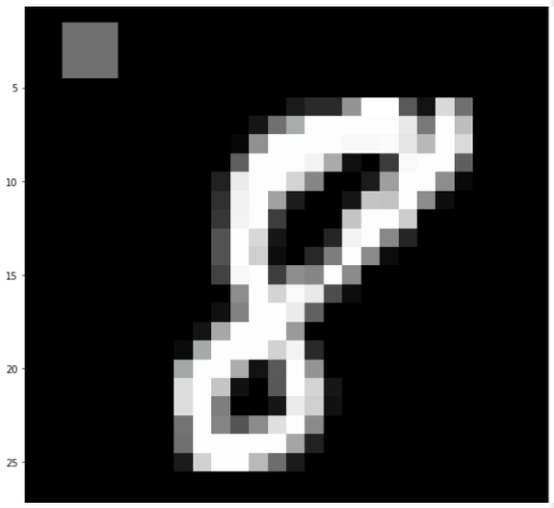
Linear Map

Concept Space

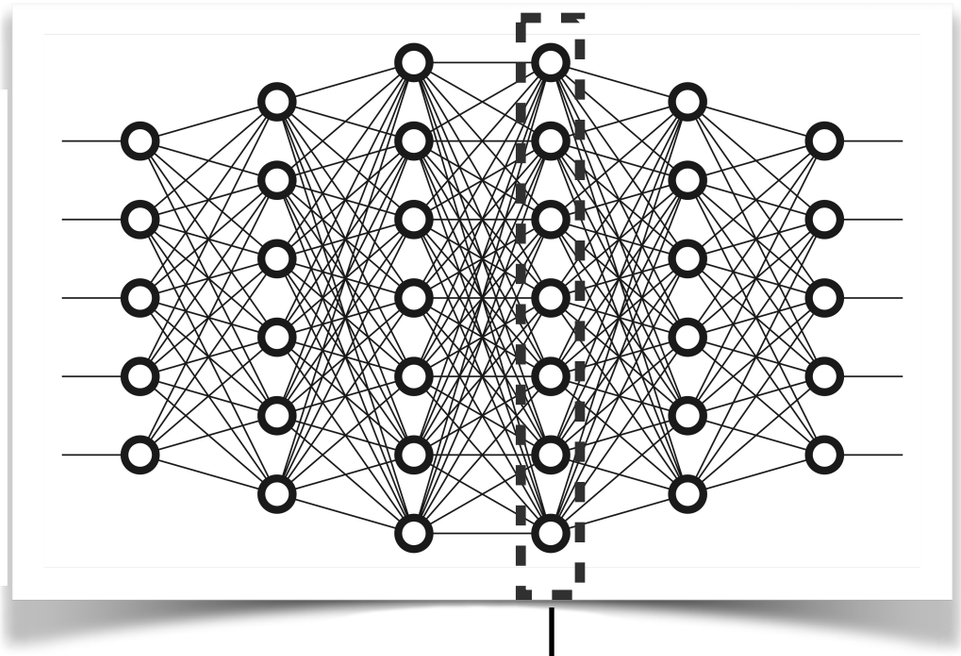
Experimenting on MNIST

- Dividing each number into 8 unique ‘strokes’ for concept space.
- Four Lines
    - Backslash, fwd slash, horizontal, vertical
- Four Curves
    - Smiley, tilted smiley, frownie, tilted frownie.
- Also adding spurious features for validation.
- Using logistic regression for linear maps and the deepest features from convolutional layers.

Concept mapping for MNIST



Input with embedded square



[0-9]

Linear Map

Concept Space

Target	Cross	Square
8	0	1

Experiment - forcing neural network to identify concepts

1. Modify MNIST digits by embedding concepts (cross, square shapes), build concept space with these concepts.

2. Train classifier.

3. Train logistic regressor to map from a hidden layer to the concept space.

4. Generate post-hoc explanations with the map created for new inputs.

Results

- A few combinations of digit class embeddings and concept spaces created accurate (>90%) mappings.

- Finding reliable maps is a hard task though.

- The tuning of the process is a potentially iterative procedure.

Future Work and Thoughts

- Need to try -
    - Automatic extraction of concepts.
    - Better ways of neuron selection.
- Concepts we perceive might not be the ones the neural net uses for inference.
- Reducing neural activity to human-readable concepts is a lossy operation even for familiar abstractions.

Thank you.