# TopoAct: Visually Exploring the Shape of Activations in Deep Learning

## Topological Data Analysis + Machine Learning

Archit Rathore, Nithin Chalapathi, Sourabh Palande
**Bei Wang***

*University of Utah
www.sci.utah.edu/~beiwang
beiwang@sci.utah.edu

https://arxiv.org/abs/1912.06332

Demo: https://github.com/architrathore/TopoAct-v2.1/
Source code: https://architrathore.github.io/TopoAct-v2.1/

The first two authors contribute equally to the work.

July 2, 2020

# Acknowledgment


- This project started with a twitter message by Chris Olah shared by Jeff Phillips.
- NSF DBI-1661375, NSF IIS-1513616, NSF IIS-1910733

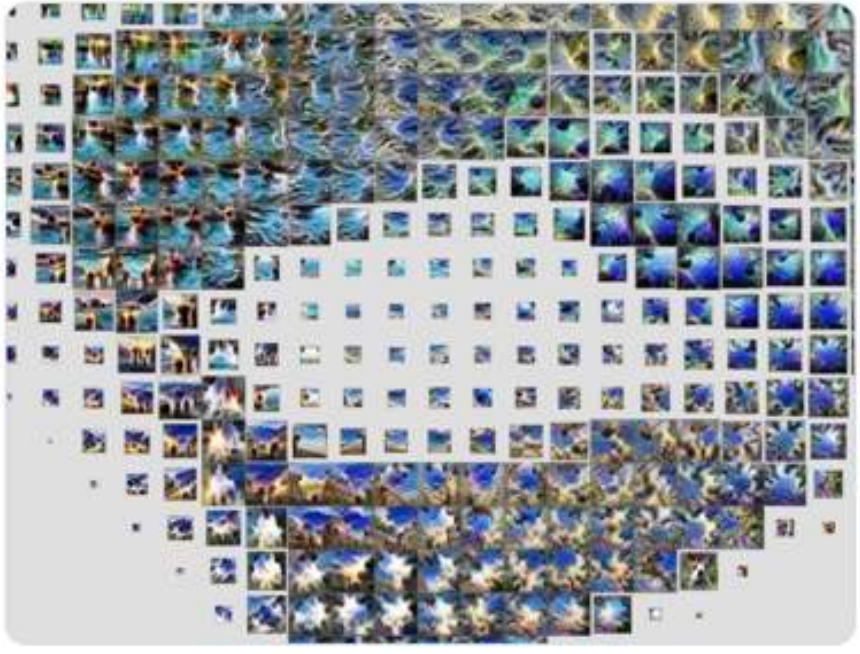
# Let us start with Twitter...

**Chris Olah**
@ch402

If you look closely at activation atlases, you sometimes see loops. For example:

Underwater → surface of water → fountain → cloudy sky → sky → underwater ??

distill.pub/2019/activatio...



10:16 AM · Mar 7, 2019 · Twitter Web Client

**Chris Olah** @ch402 · Mar 7, 2019
Replying to @ch402
Another one?

Ground → grass → leaves → flowers → feathers → birds → bird legs → legs → ground.

distill.pub/2019/activatio...

2 2 11

**Chris Olah** @ch402 · Mar 7, 2019
One could argue for these loops being genuine topological features of the underlying ImageNet data, since neural nets are a continuous map and t-sne/umap try to preserve neighborhood structure.

(Counter argument: be cautious to draw conclusions from t-sne/umap layouts!)

2 4 19

# Let us start with Twitter...

TDA + NN Representations?

# Interpretability: a main challenge in deep learning

What representations have these neural networks learned that could be made human interpretable?

- Given a trained NN, we probe neuron activations (combinations of neuron firings) in response to a particular input image.
- With millions of input images, we would like to obtain a global view of what the neurons have learned by studying neuron activations at a particular layer, and across multiple layers of the network.

# Topology of neuron activations

- What is the shape of the space of activations?
- **What is the organizational principle behind neuron activations?**
- How are the activations related within a layer and across layers?

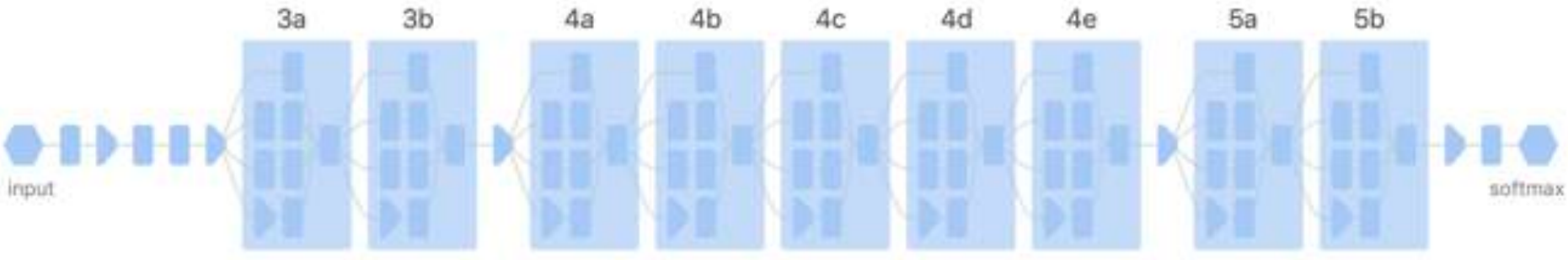
Ingredients:

- Neuron activation vectors as point clouds
- *Mapper graphs as summary graphs*
- Feature visualization
- Interactive and exploratory visual analytics

# Take home message for TopoAct

- Capture topological structures (branching and loop structures) in the space of activations that are hard to detect via DR
- Offer new perspectives on how a NN "sees" the input images.

# GoogLeNet (InceptionV1)




Trained on ImageNet ILSVRC.

Image: `https://distill.pub/2019/activation-atlas/`

# What is neuron activation?

Fix a pre-trained model, a particular layer of interest, an input image:

- We feed an input image to the network and collect the activations (the numerical values of how much each neuron has fired with respect to the input). The activation of a neuron is a non-linear transformation (i.e., a function) of its input.
- A single neuron produces a collection of activations from a number of overlapping patches of an input image.
- We randomly sample a single activation from these patches.

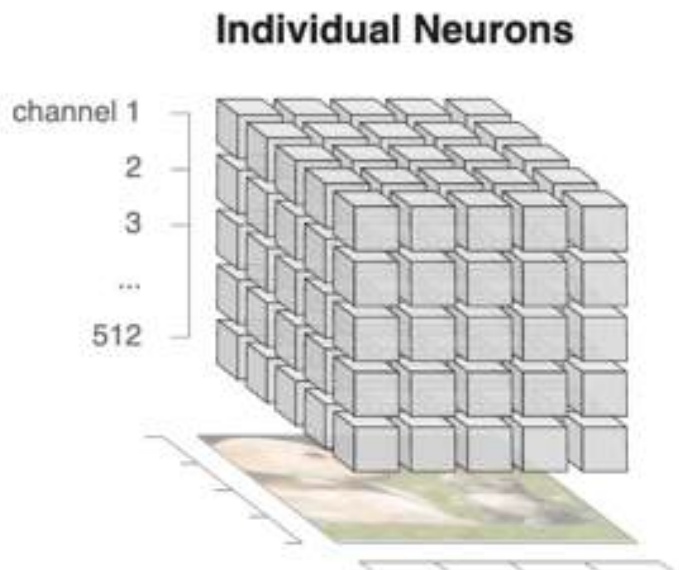



Image: https://distill.pub/2018/building-blocks/

# What is neuron activation?

Fix a pre-trained model, a particular layer of interest, an input image:

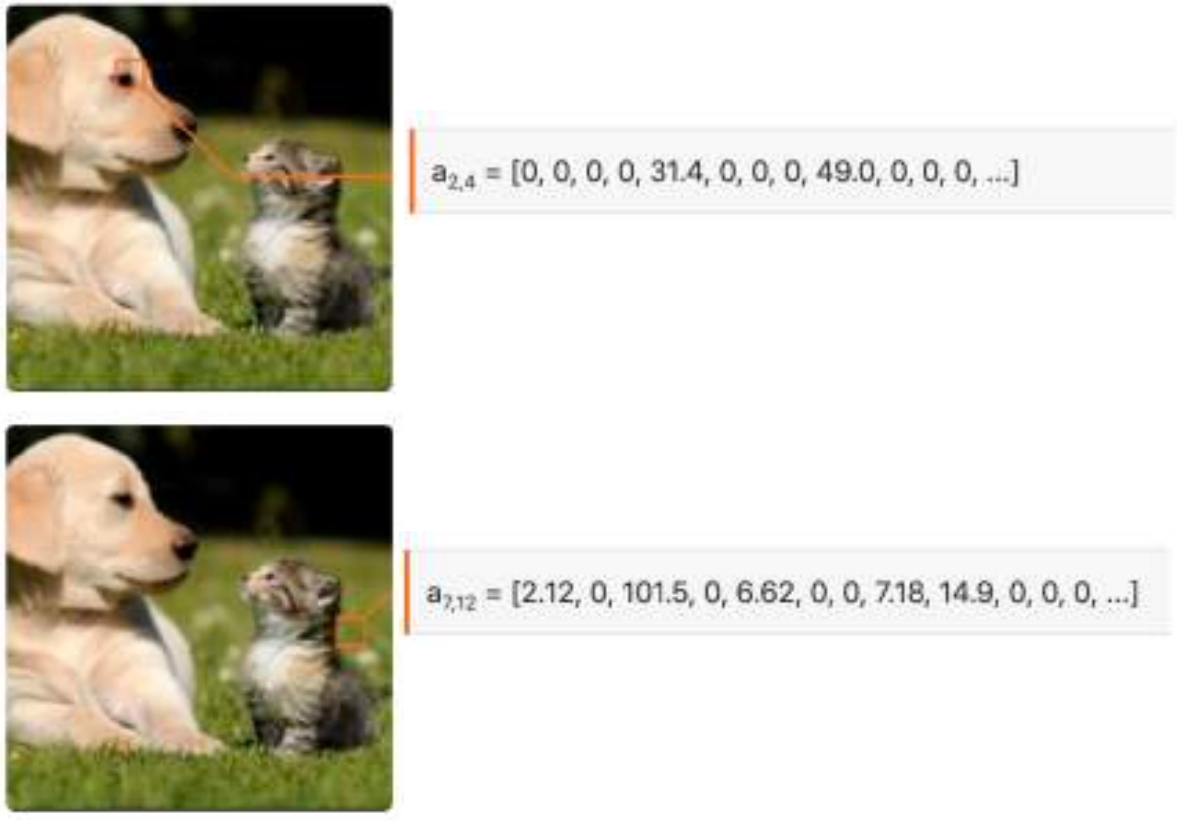
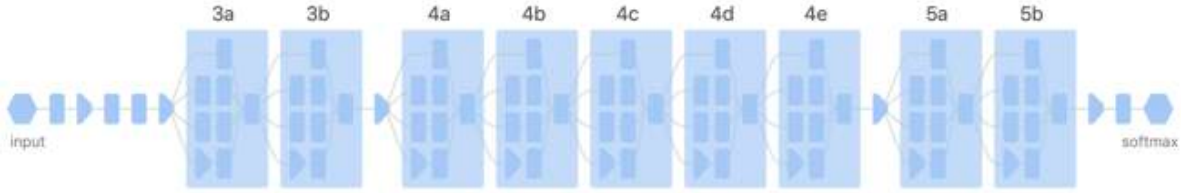


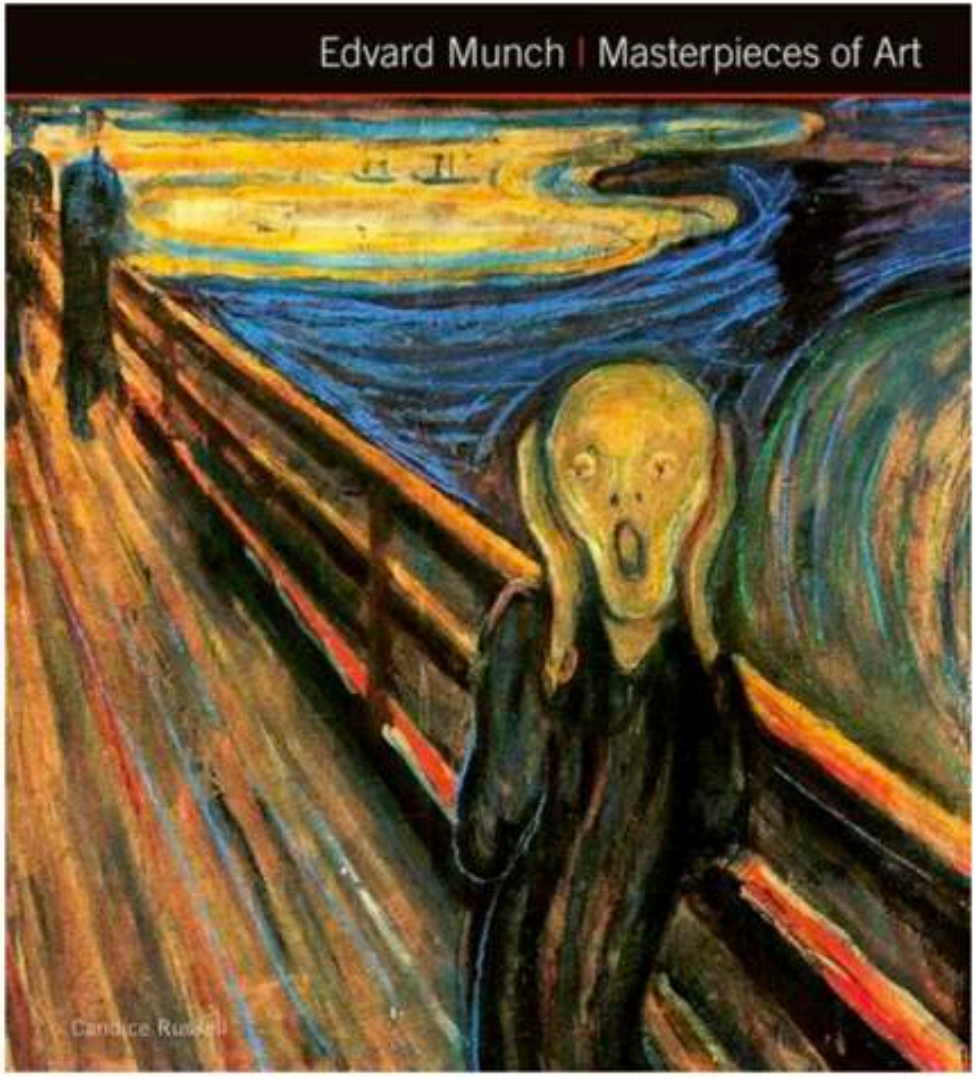

Image: https://distill.pub/2018/building-blocks/

# TDA of activation vectors for InceptionV1



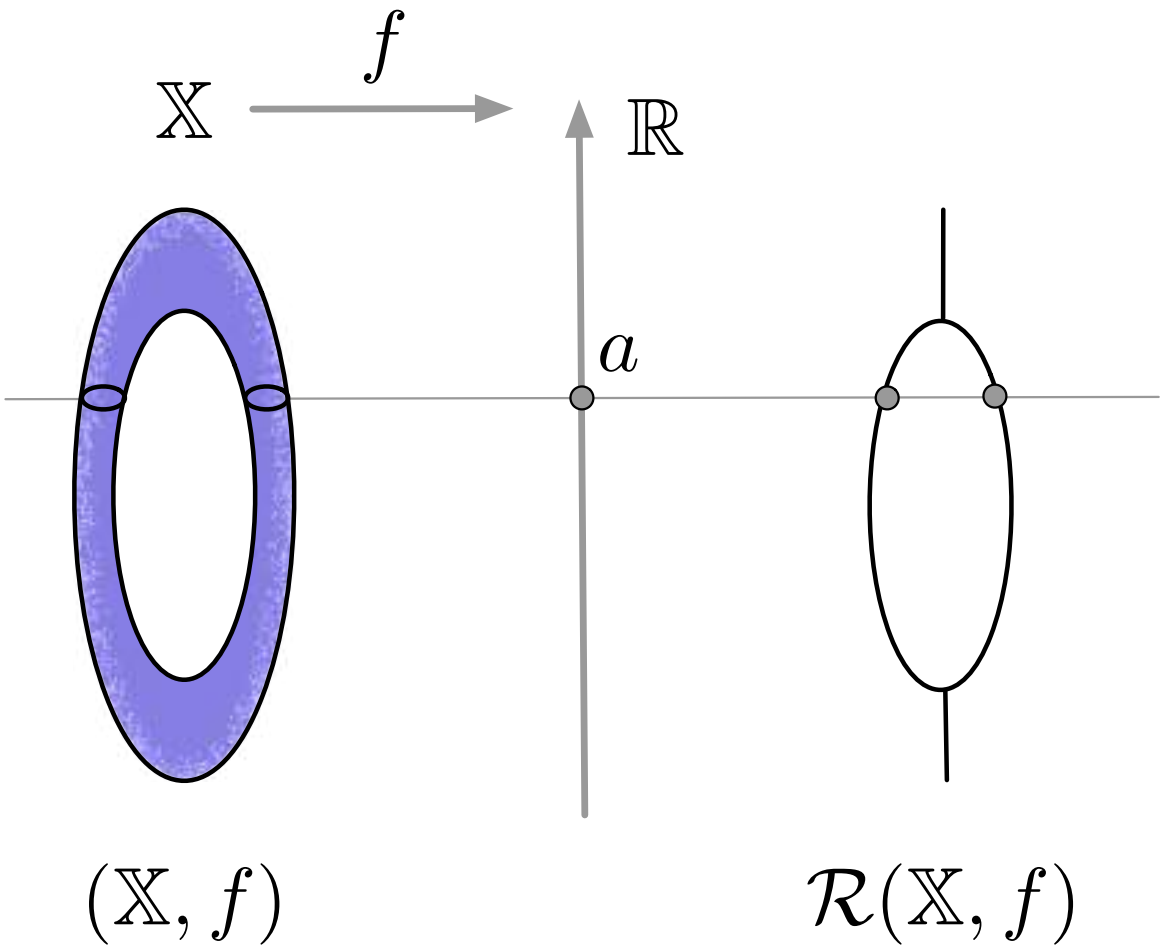

- Suppose an input image has $14 \times 14$ patches.
- A neuron within layer 4c outputs $14 \times 14$ activations per image.
- Randomly sample a single activation from the $14 \times 14$ patches.
- Each activation vector is high-dimensional; its dimension depends on the number of neurons in that layer.
- Layers 3a, 3b, and 4a have 256, 480, and 512 neurons respectively, producing point clouds in 256, 480, and 512 dimensions.
- $300,000$ images $\rightarrow$ $300,000$ activation vectors for a given layer.
- We then apply the mapper framework to obtain topological summary graphs of these point clouds.

# Here comes the math...



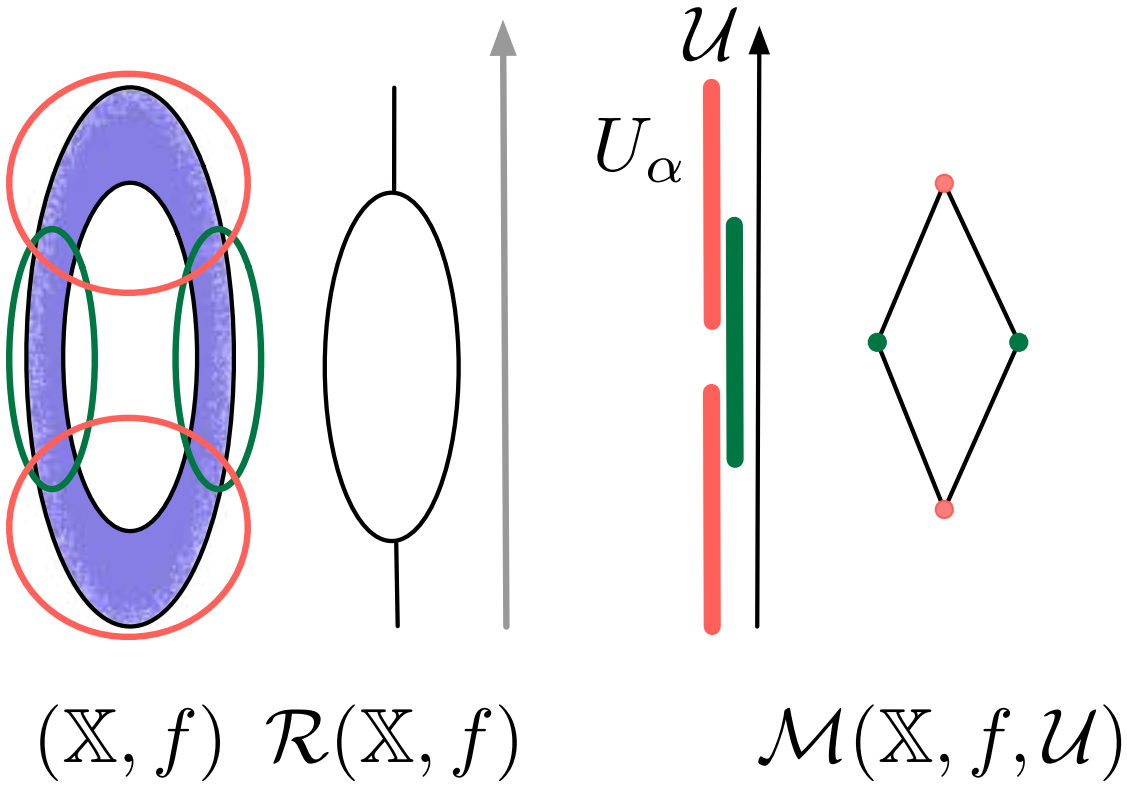
# A tale of two topological constructs: Reeb graph

Reeb graph $\mathcal{R}(\mathbb{X}, f)$ encodes the connected components of the level sets $f^{-1}(a)$ for $a$ ranging over $\mathbb{R}$.



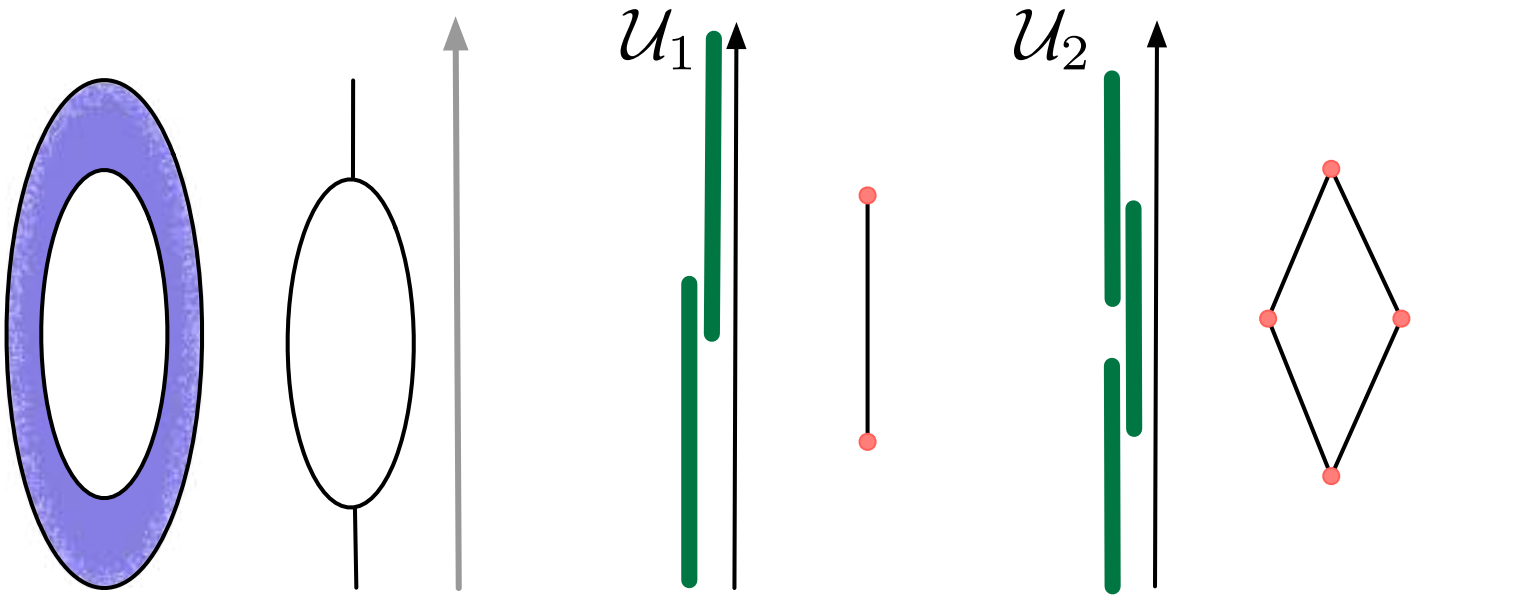
# A tale of two topological constructs: mapper graph

Given a finite good cover $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ of $f(\mathbb{X})$, let $f^*(\mathcal{U})$ denote the cover of $\mathbb{X}$ obtained by considering the path-connected components of $f^{-1}(U_\alpha)$ for each $\alpha$. The mapper construction of $(\mathbb{X}, f)$ is defined to be the nerve of $f^*(\mathcal{U})$, denoted as $\mathcal{M}(\mathbb{X}, f, \mathcal{U})$



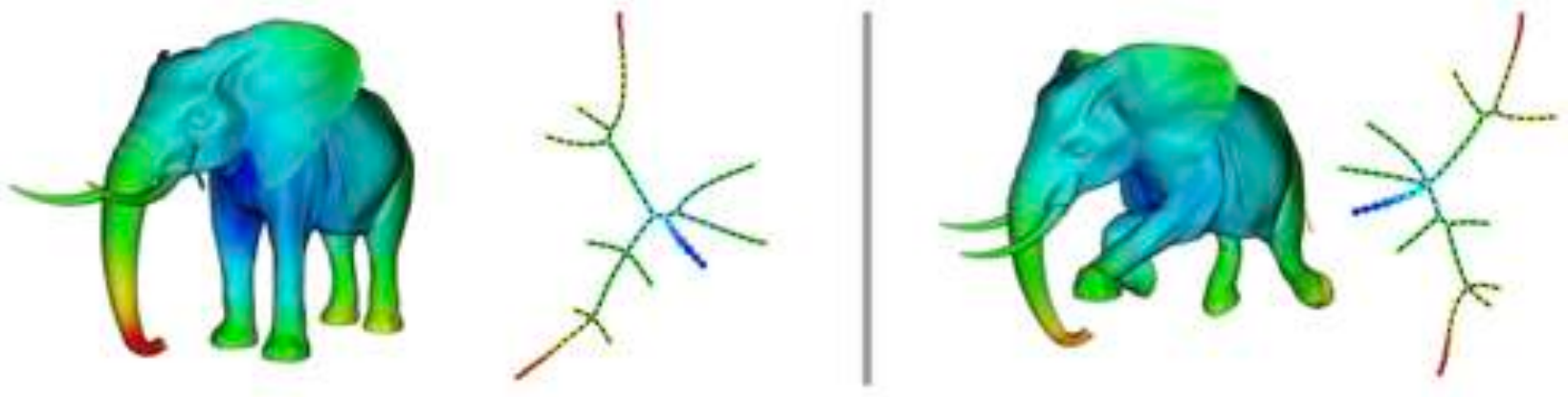

$\mathbb{X}$: a manifold or a point cloud sample

# Mapper graphs at different resolution



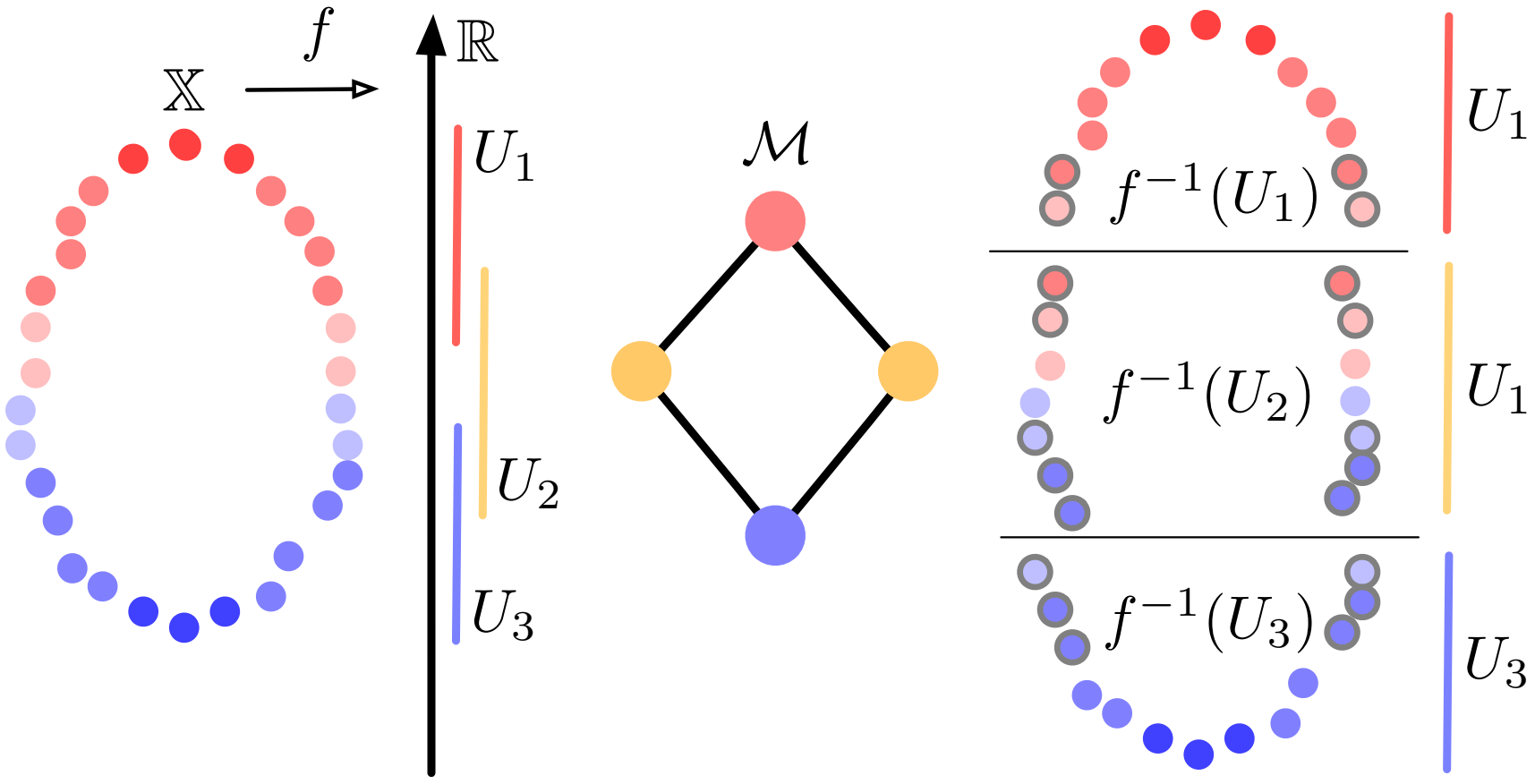
# Mapper in TDA

The mapper construction is widely appreciated by the practitioners...



Singh et al. (2007)

# Mapper graphs as summary graphs for point cloud data



# Feature visualization by optimization

**Dataset Examples** show us what neurons respond to in practice

**Optimization** isolates the causes of behavior from mere correlations. A neuron may not be detecting what you initially thought.

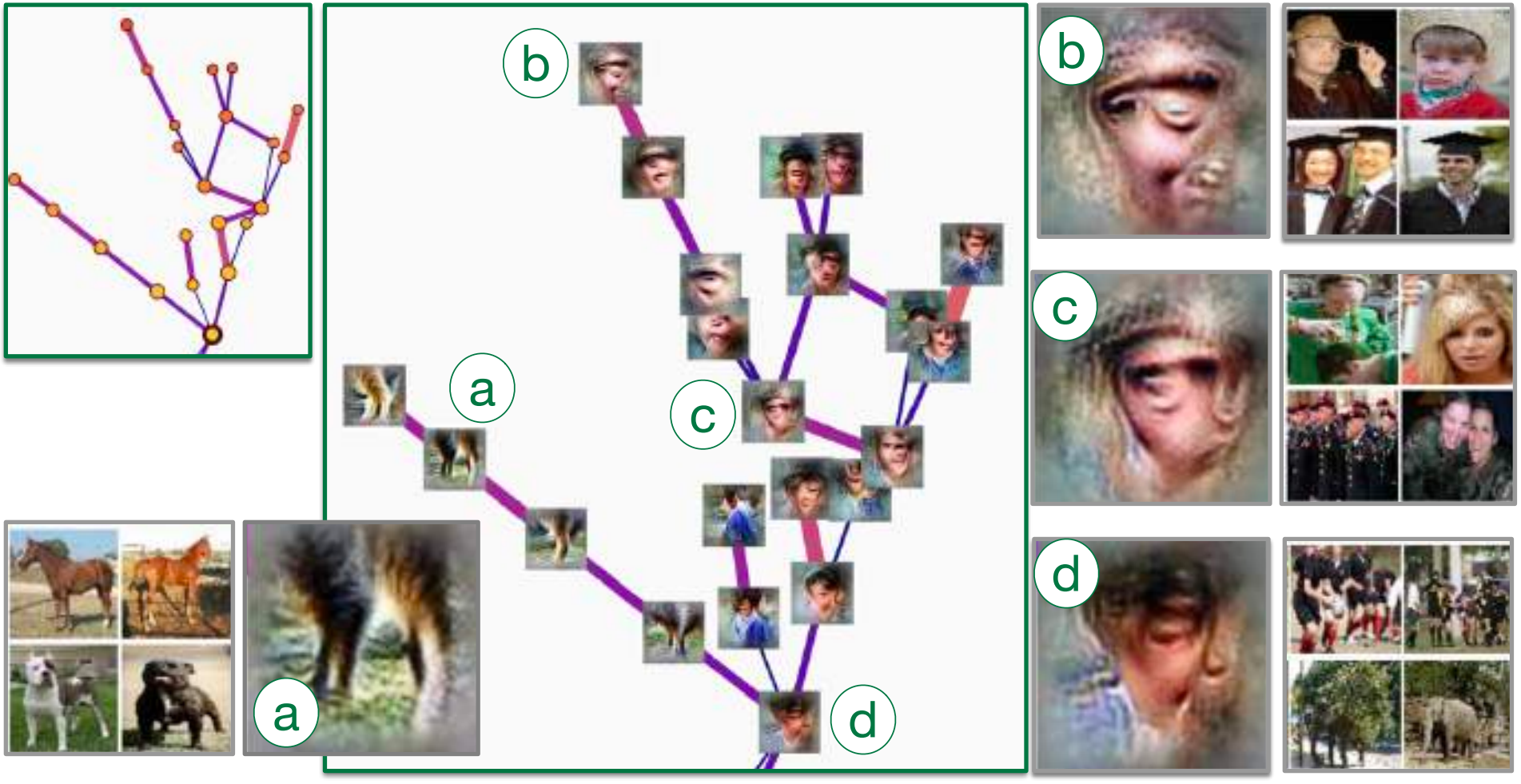
Baseball—or stripes?
*mixed4a, Unit 6*

Animal faces—or snouts?
*mixed4a, Unit 240*

Clouds—or fluffiness?
*mixed4a, Unit 453*

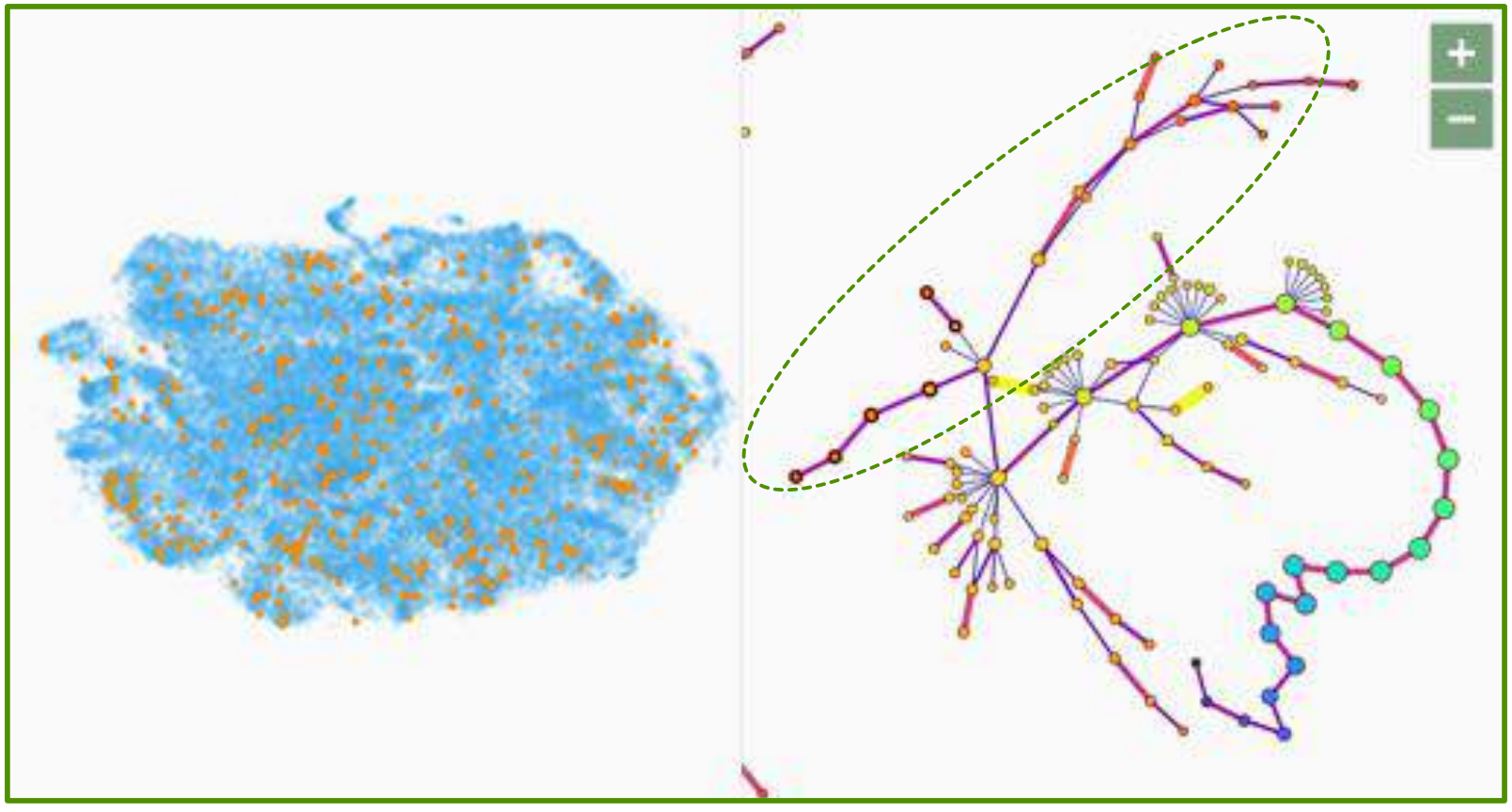
Image: https://distill.pub/2017/feature-visualization/

# Results: leg-face bifurcation (branching)



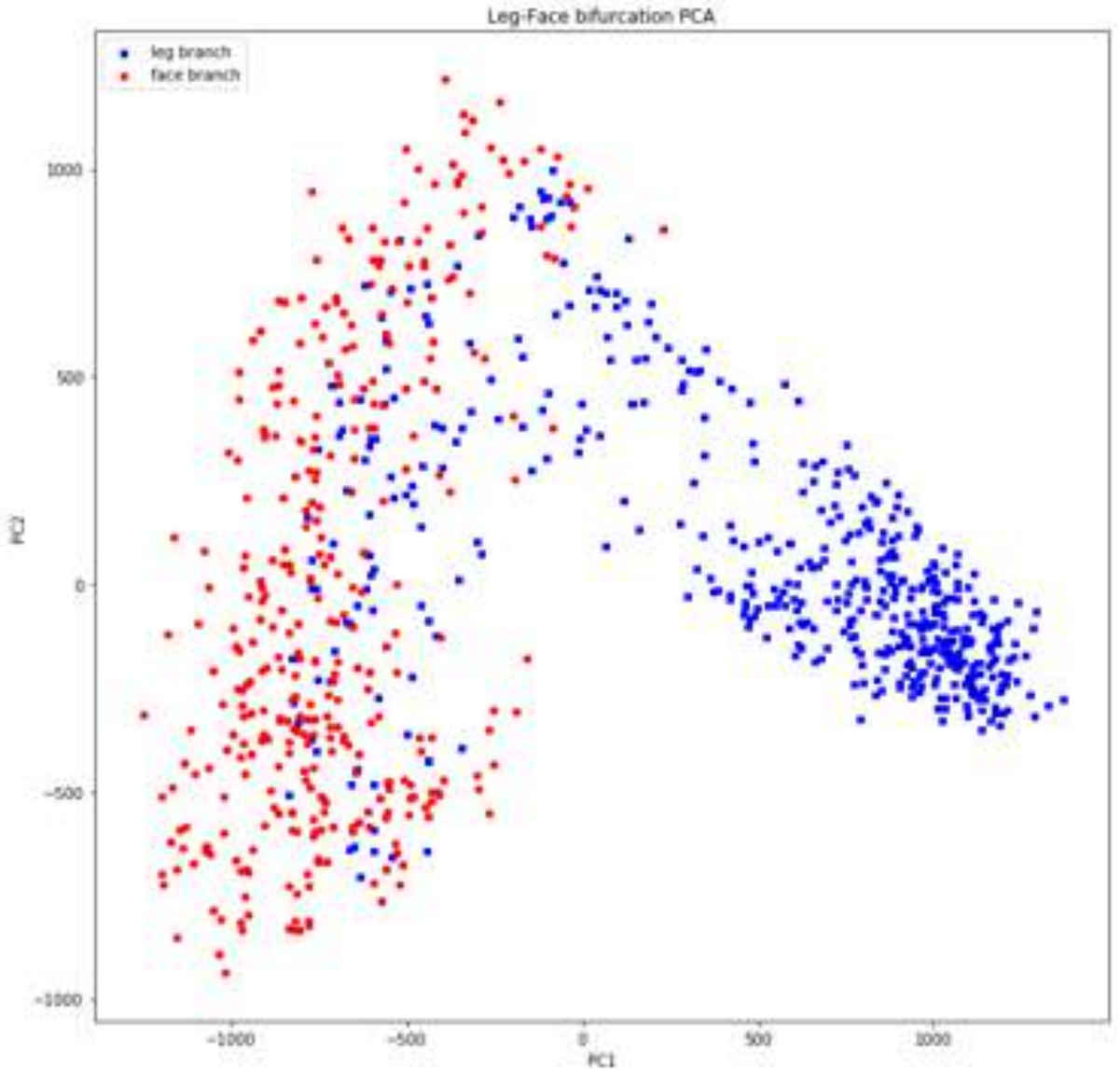

overlap-30-eps-4c

# Comparison with t-SNE



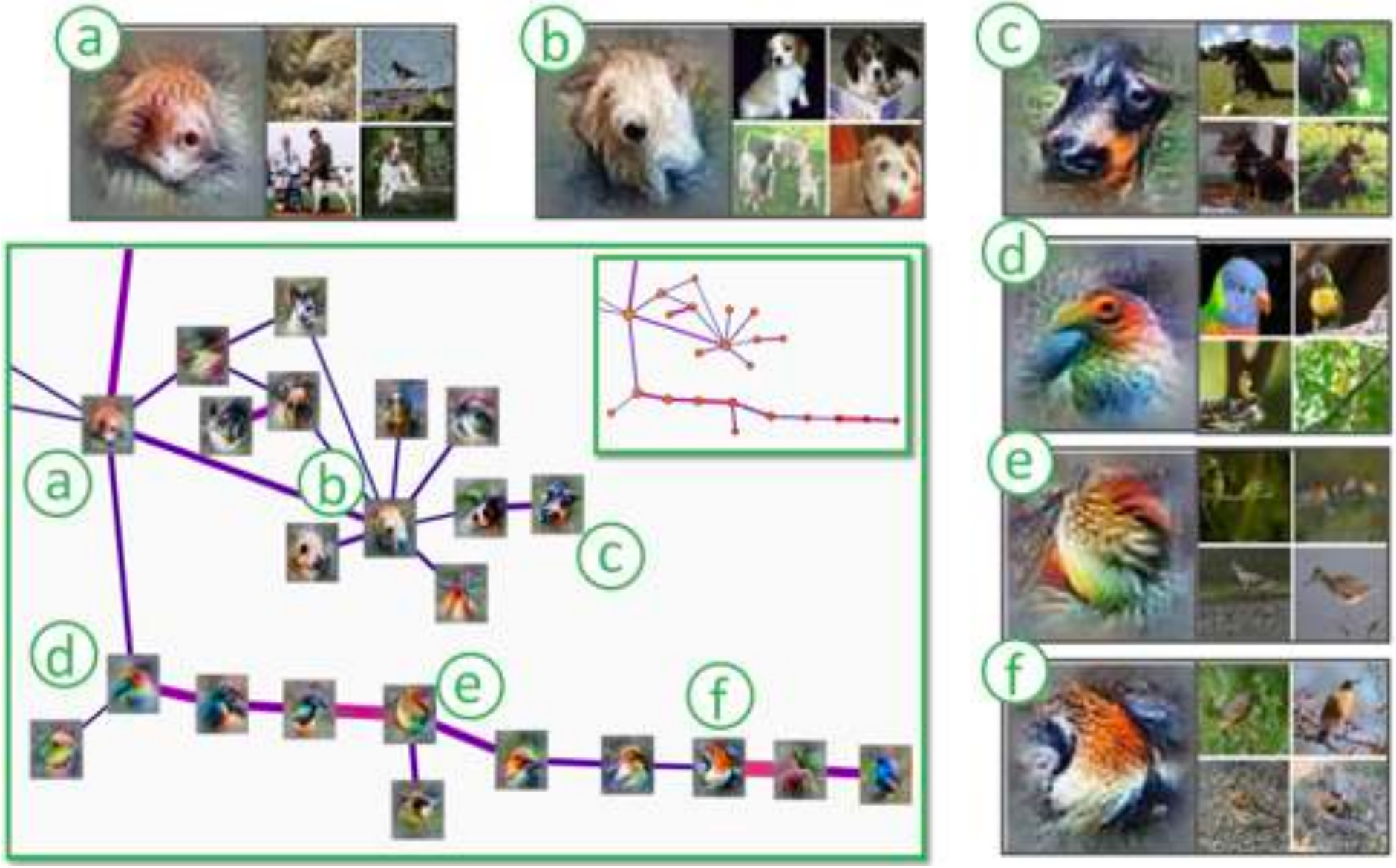
Highlighting activation vectors that belong to the leg-face bifurcation as orange points in the t-SNE projection.

# PCA of the leg-face bifurcation



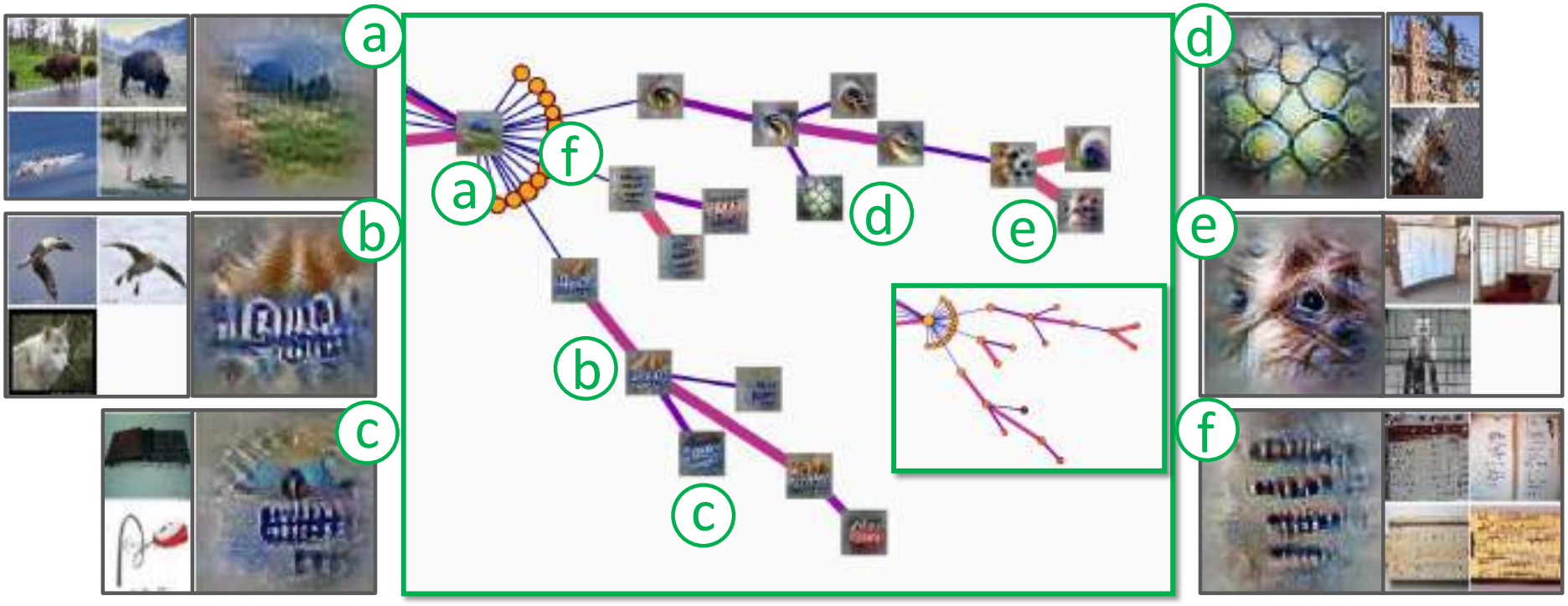

Refined analysis (and validation) of bifurcations

# Results: bird-mammal bifurcation



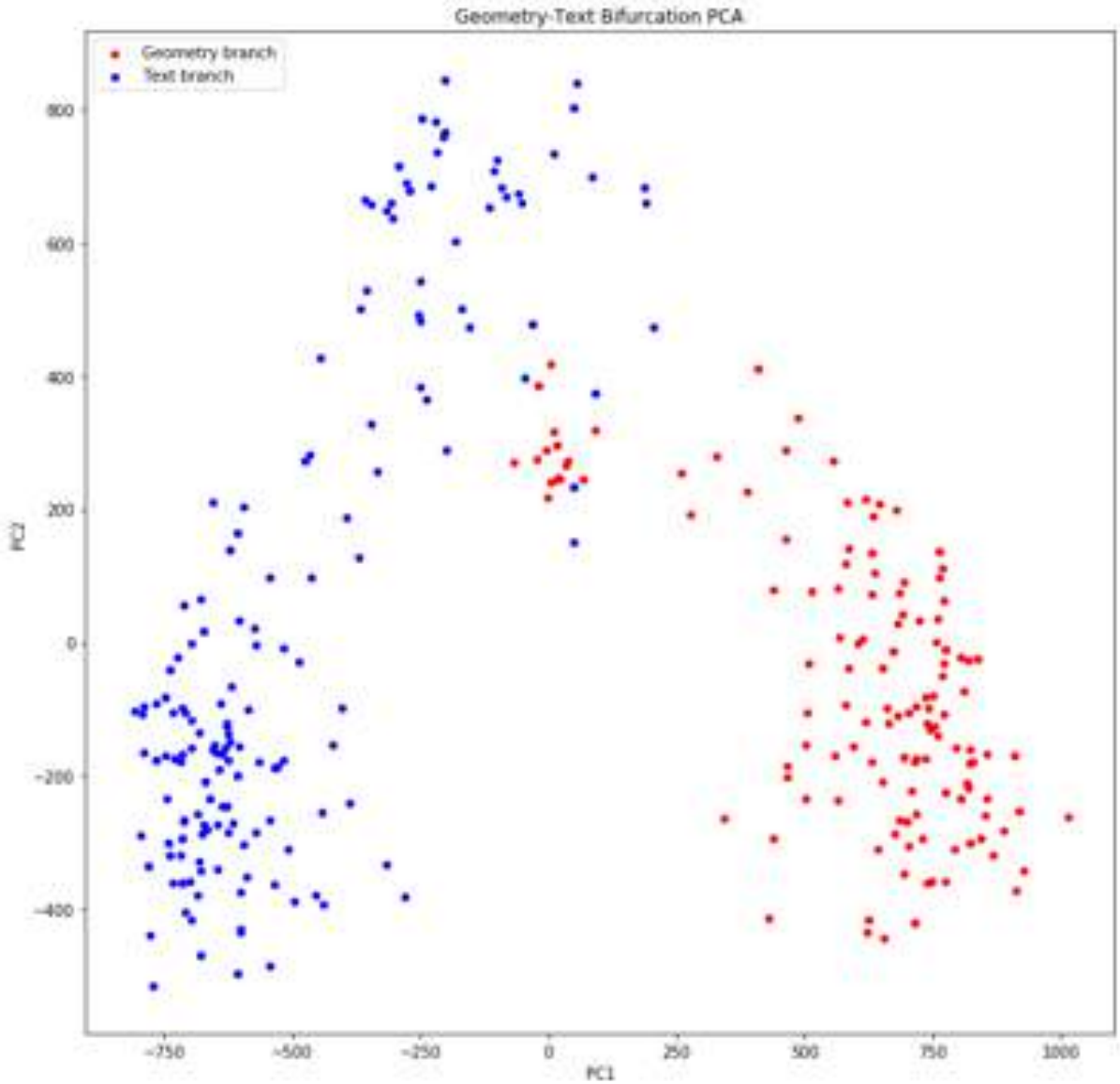

overlap-30-5a

# Results: geometry-text bifurcation



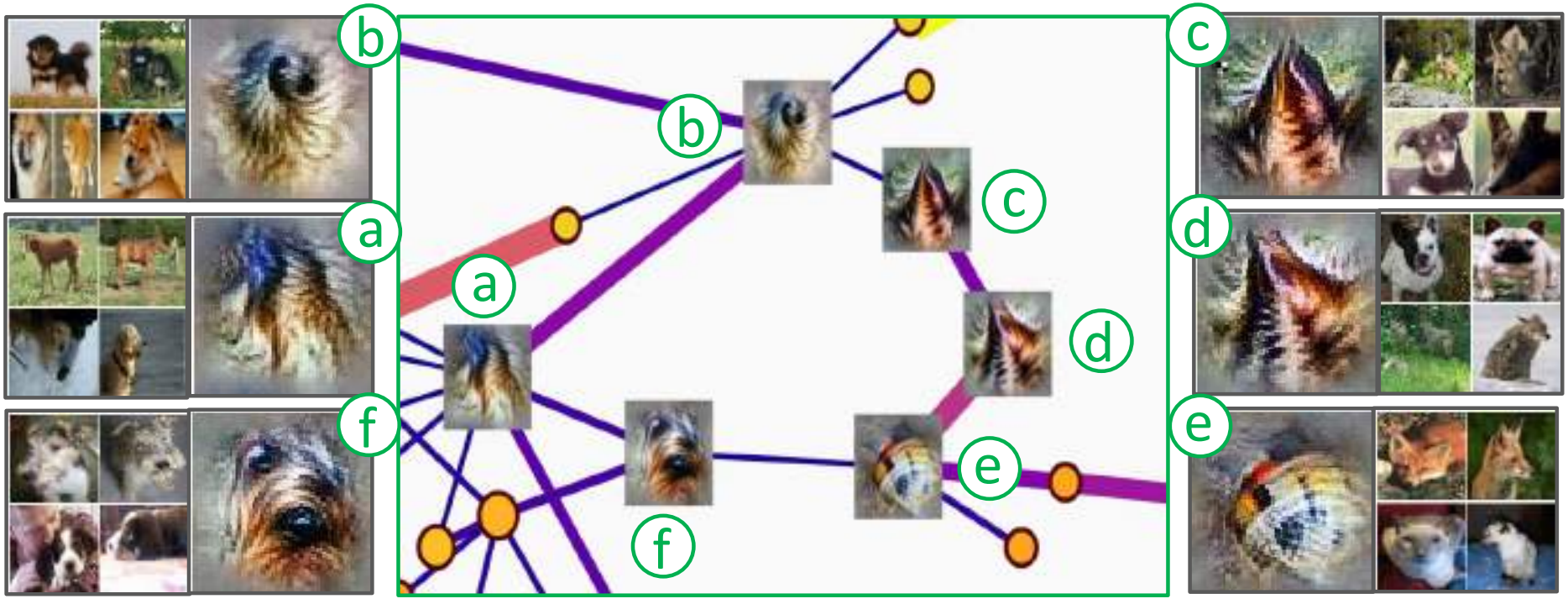

overlap-30-4b

# PCA of the geometry-text bifurcation



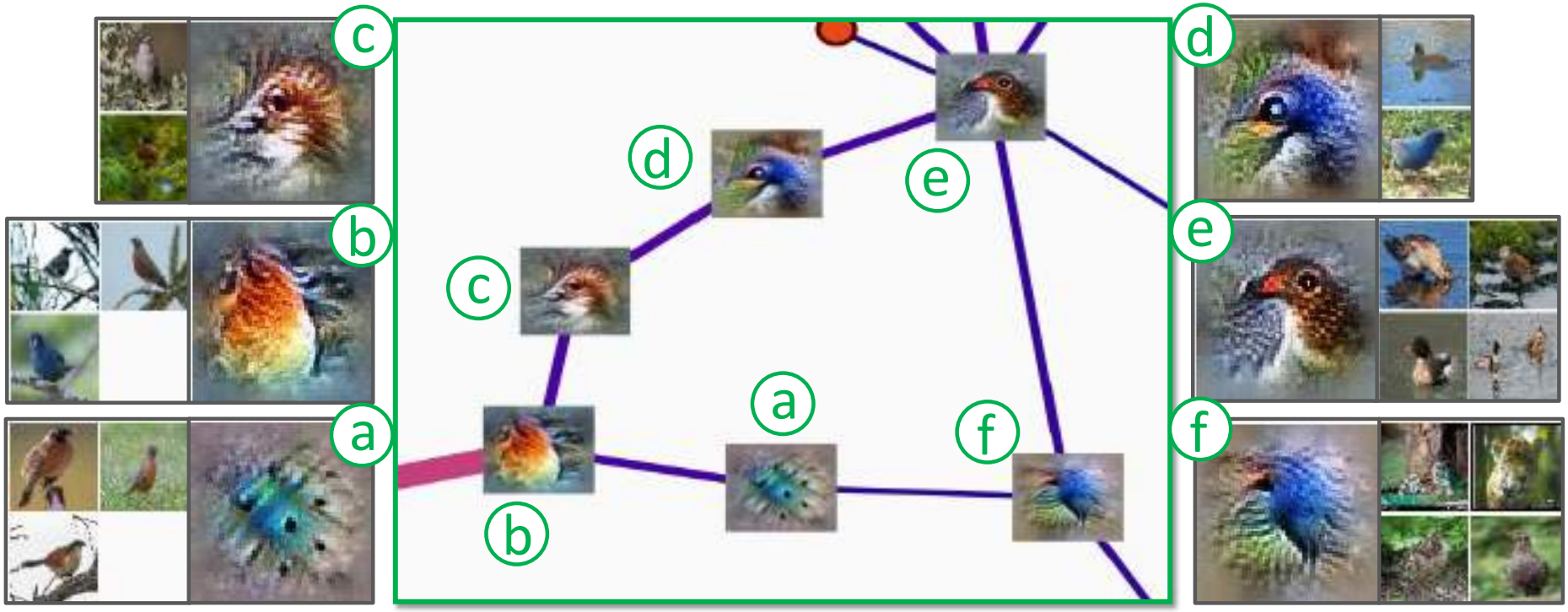

Refined analysis (and validation) of bifurcations

# Results: fur-nose-ear-eye loop



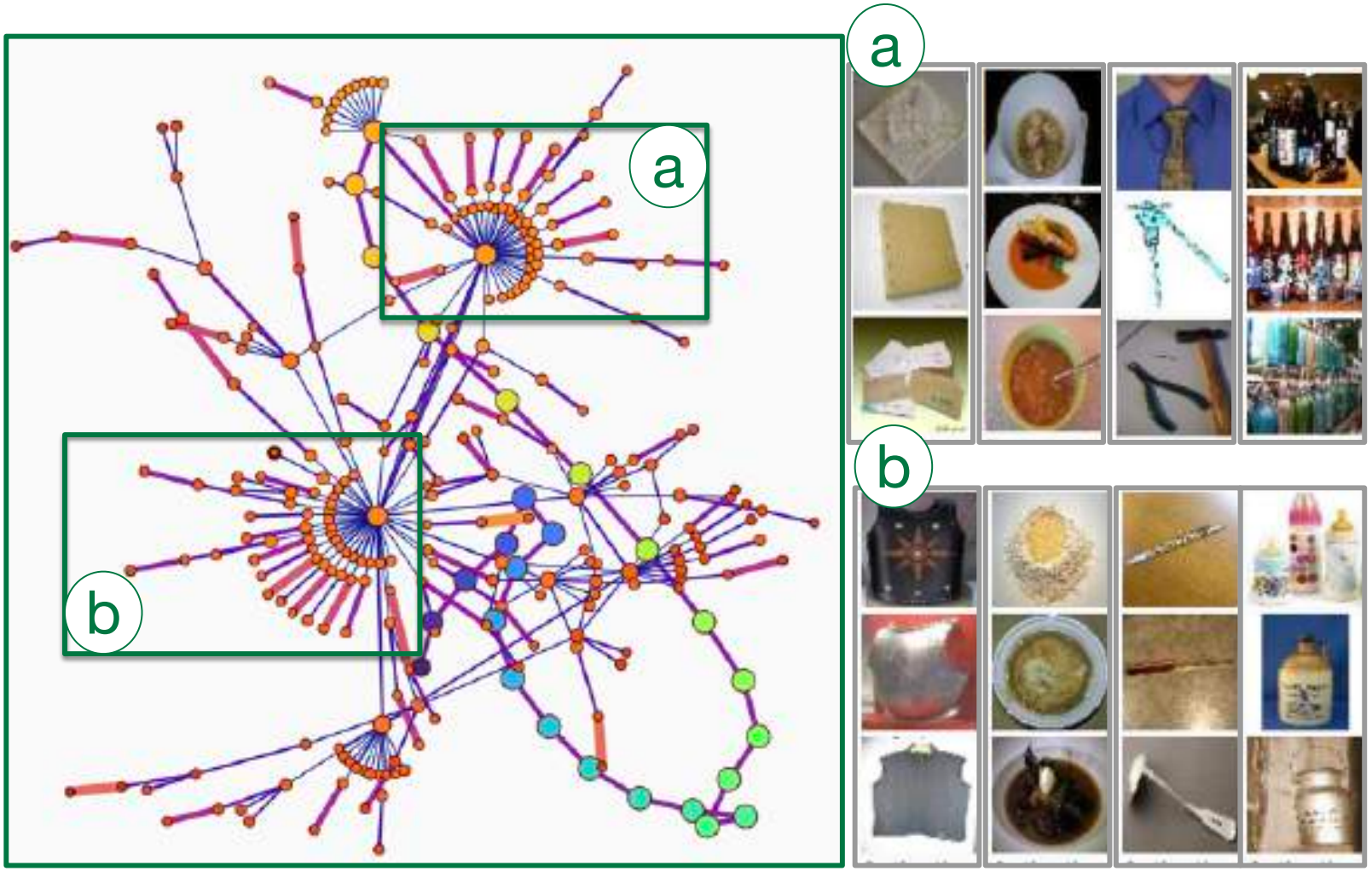

overlap-30-4d

# Results: face-body-leg loop of birds



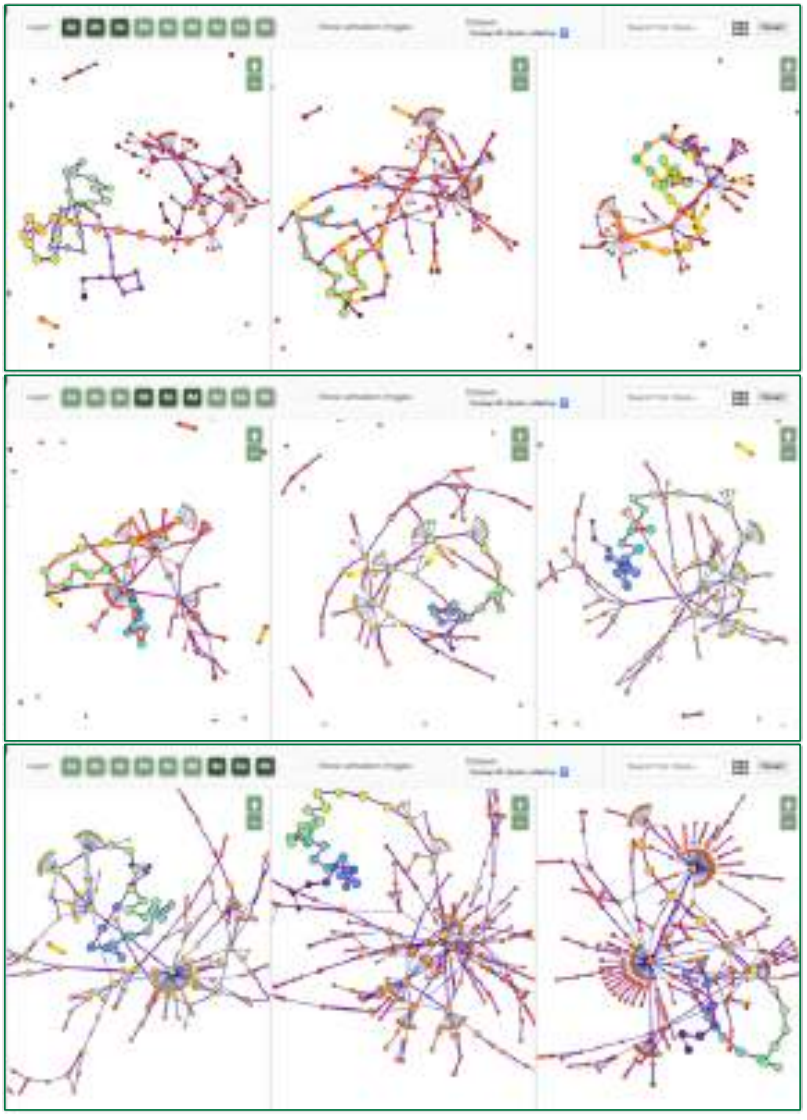

overlap-30-eps-5a

# Results: distribution of branching structures




5b-overlap-30-epsilon-adaptive

# Results: multilayer summary graphs



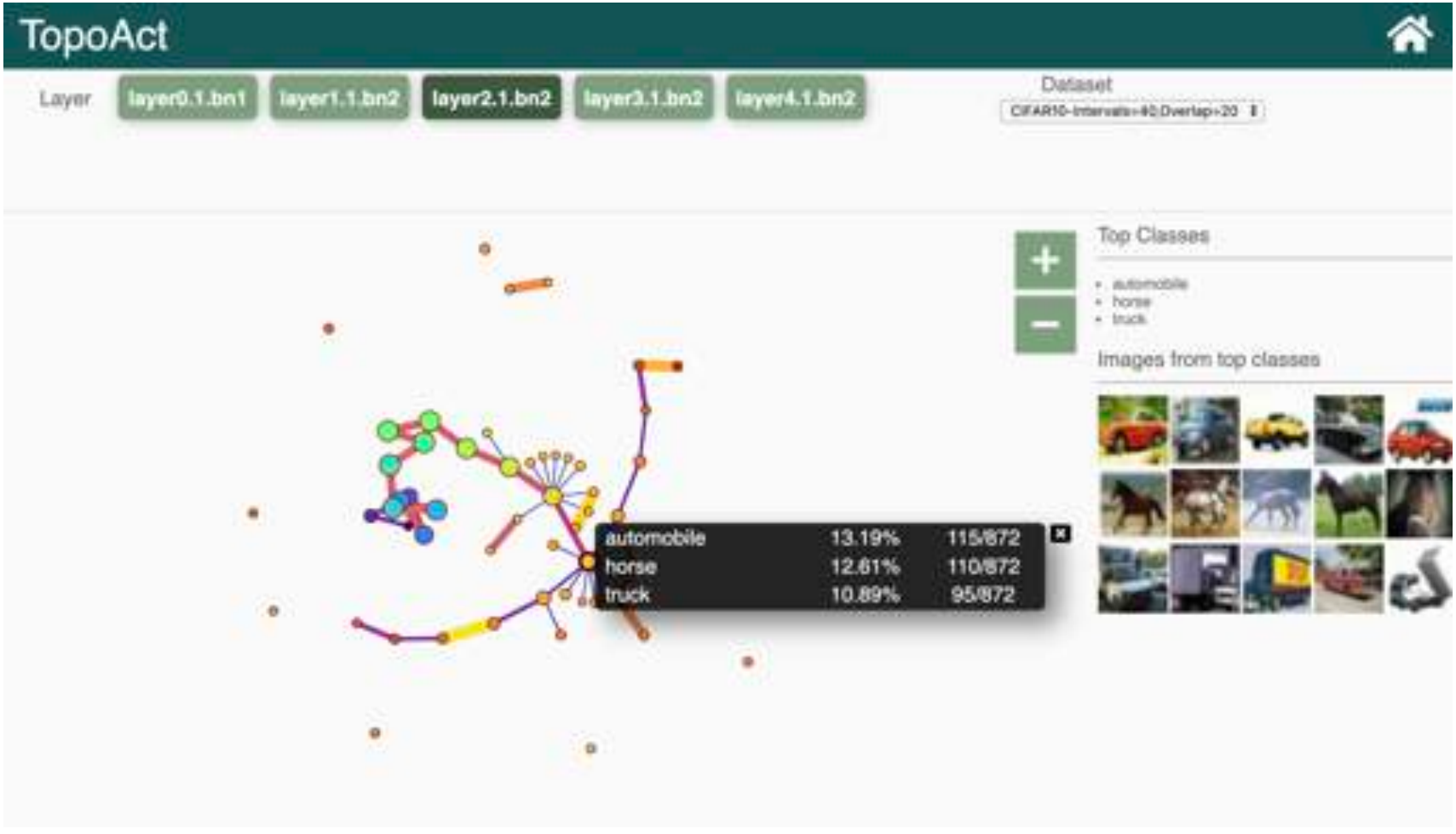
# Code and Demo

https://github.com/architrathore/TopoAct-v2.1/
https://architrathore.github.io/TopoAct-v2.1/

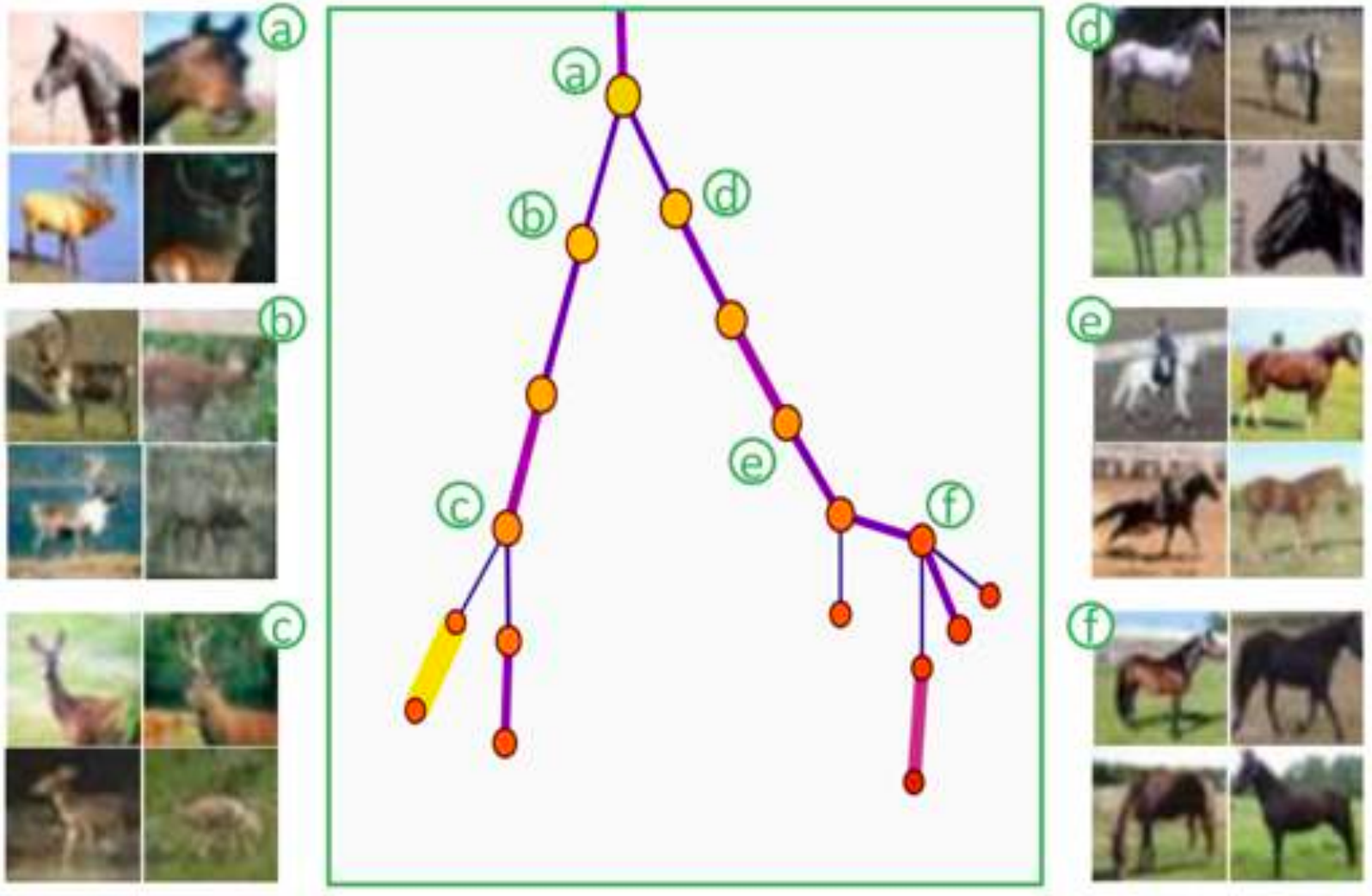
# Discussions

- Generality: other architecture, other datasets
- Parameter tuning
- Scalability
- Stability
- $L_2$ Norm and Adaptive Cover

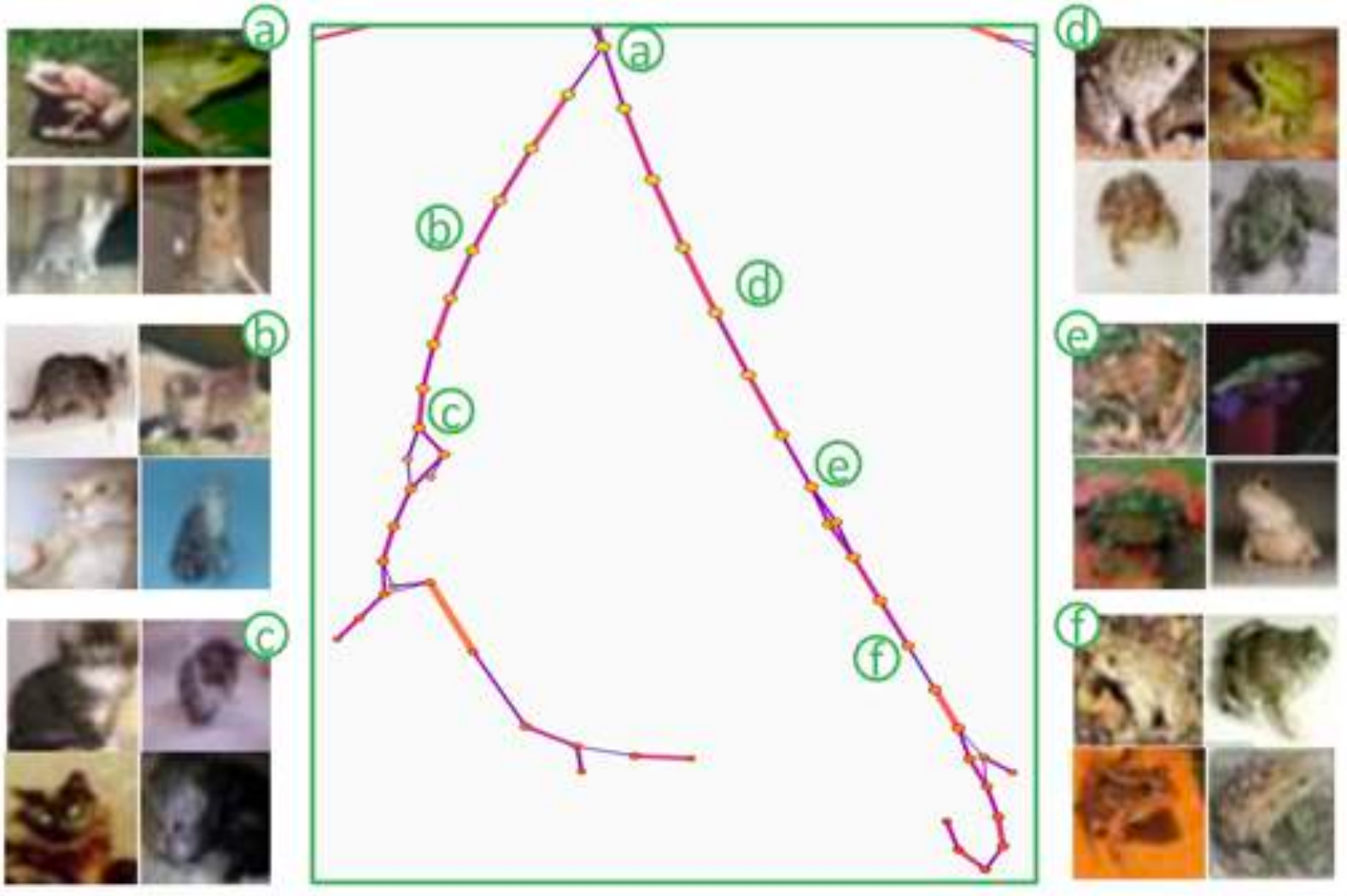
# Applying TopoAct to ResNet trained on CIFAR



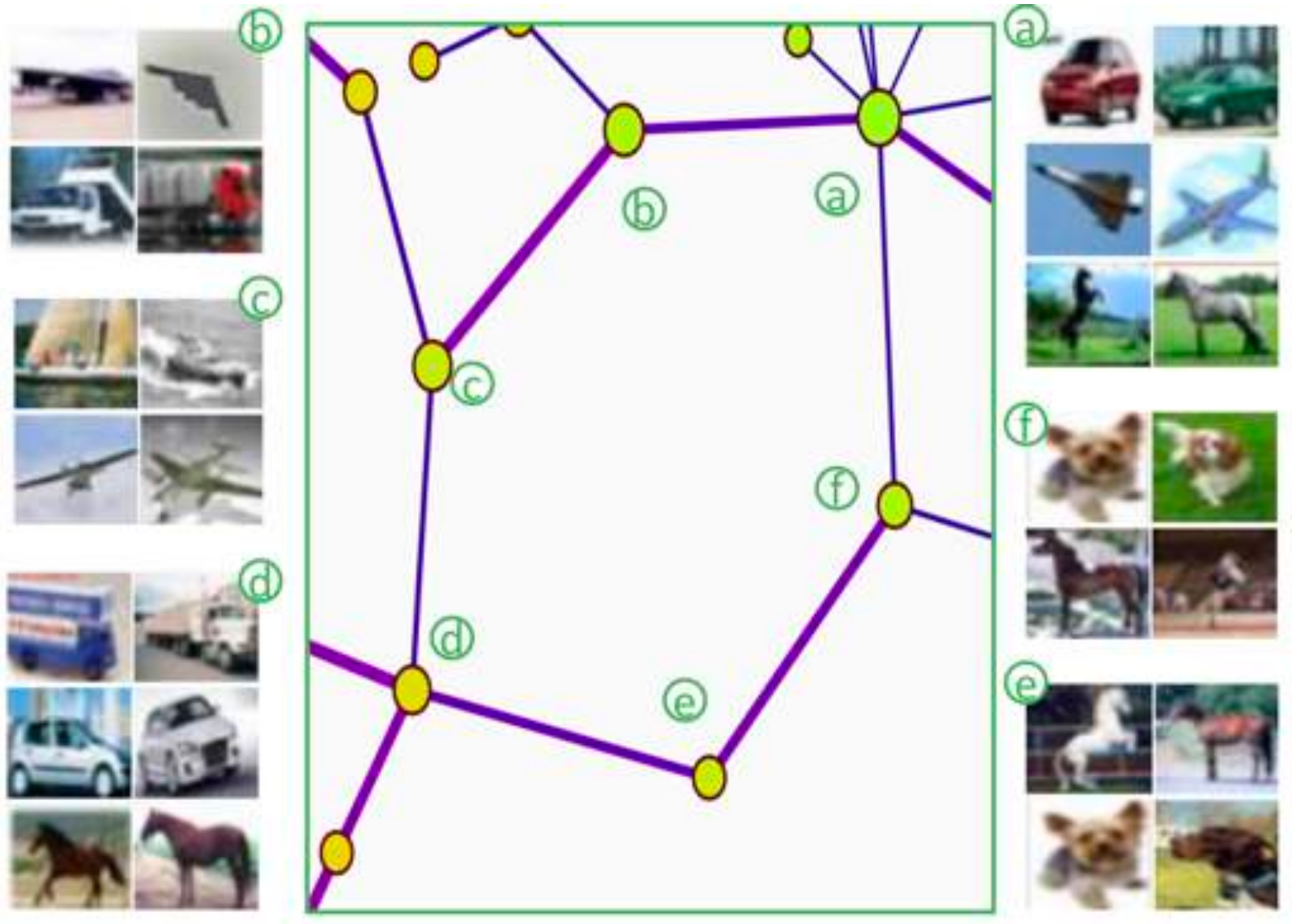
# Results CIFAR: horse-deer bifurcation



# Results CIFAR: frog-cat bifurcation



# Results CIFAR: Airplane-ship-horse-dog-truck loop



# Discussions

- Generality: other architecture, other datasets
- Parameter tuning
- Scalability
- Stability
- $L_2$ Norm and Adaptive Cover

# References I

Singh, G., Mémoli, F., and Carlsson, G. (2007). Topological methods for the analysis of high dimensional data sets and 3D object recognition. In *Eurographics Symposium on Point-Based Graphics*, pages 91–100.