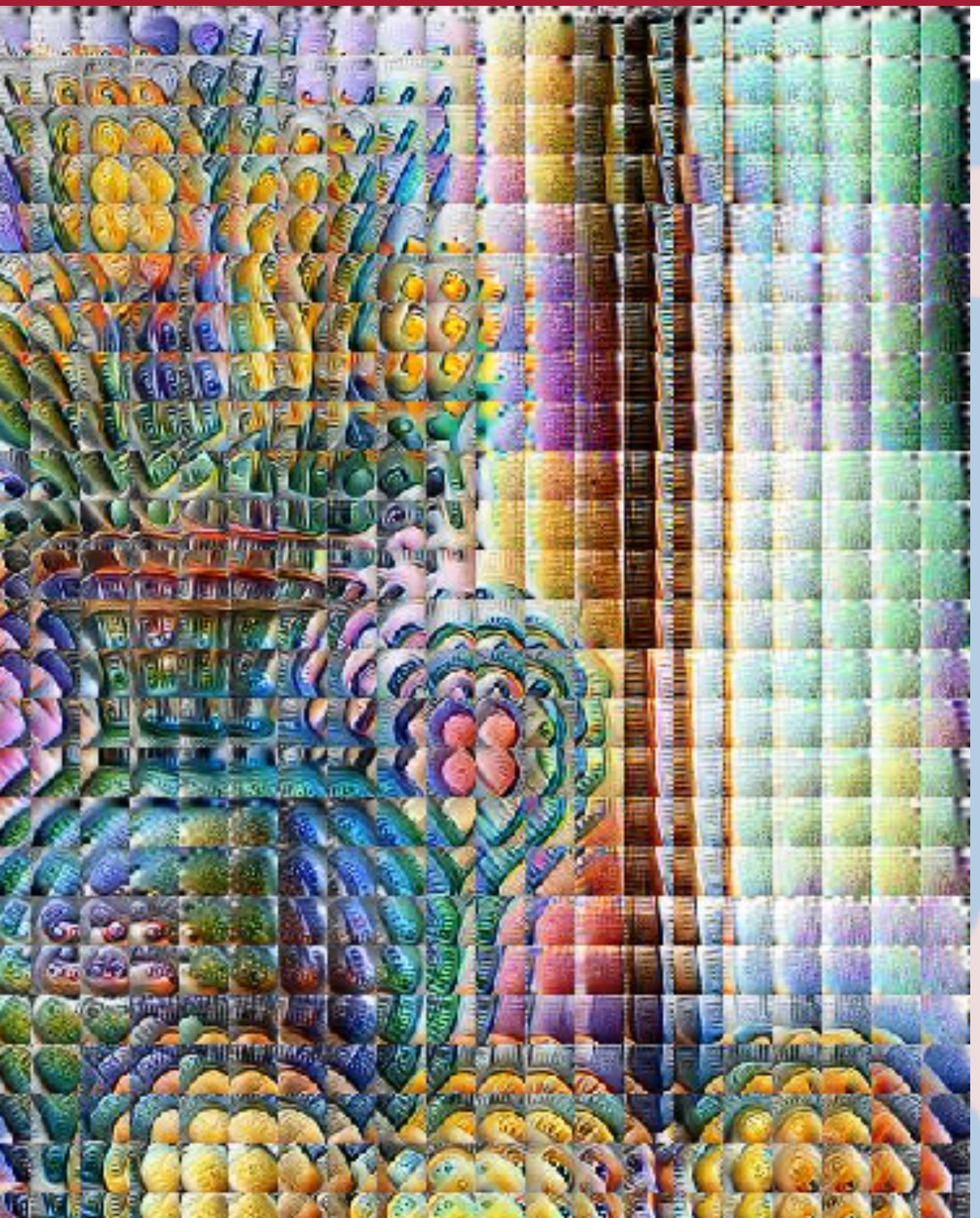
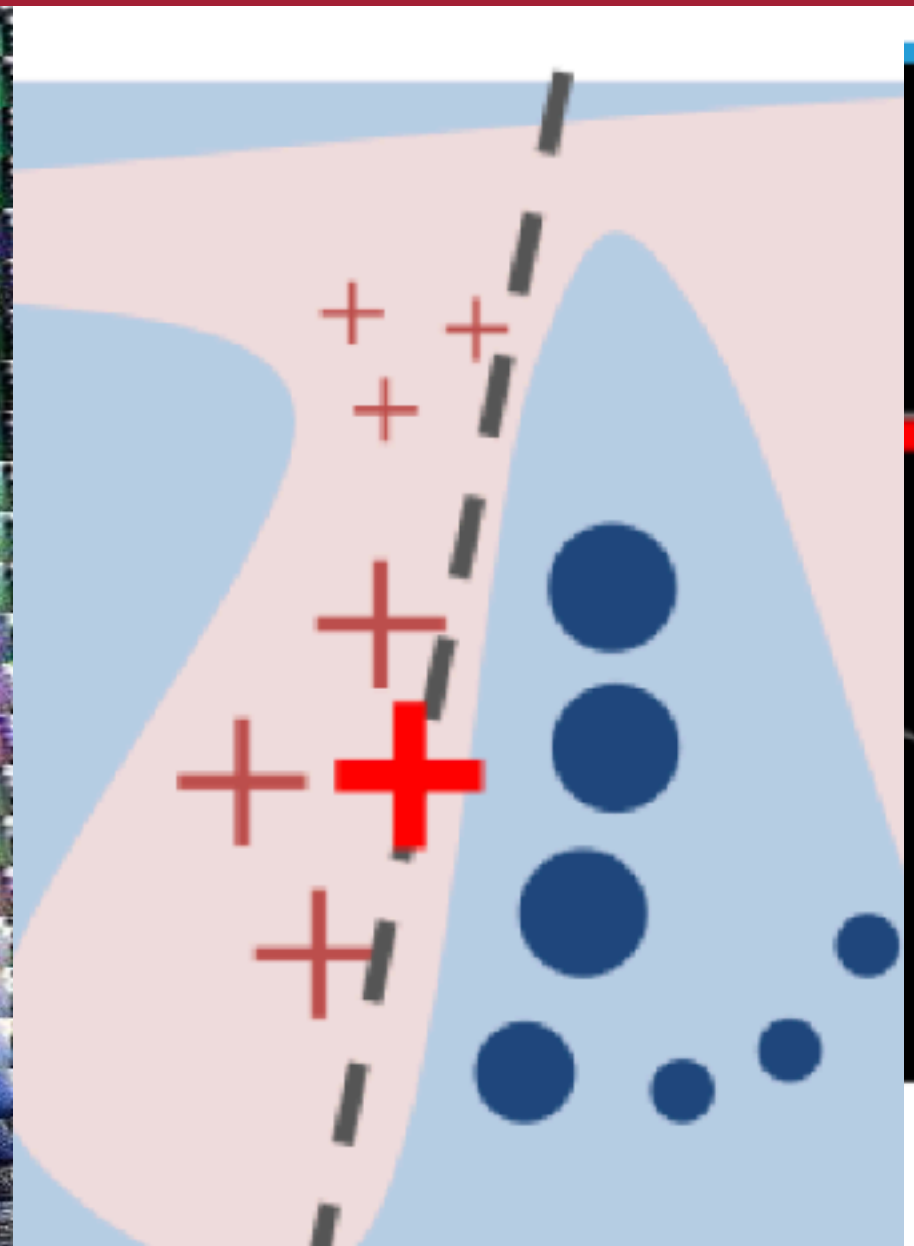
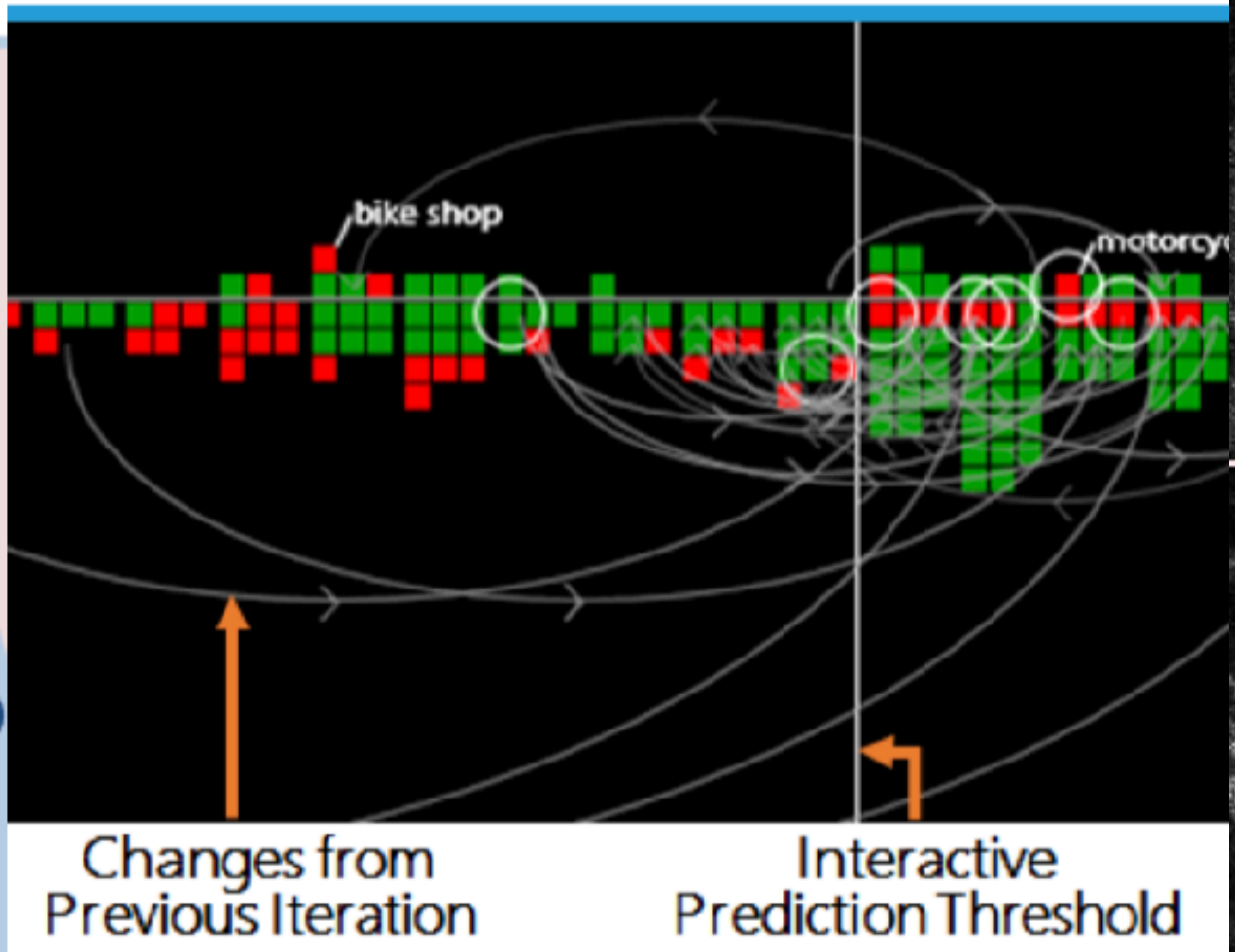
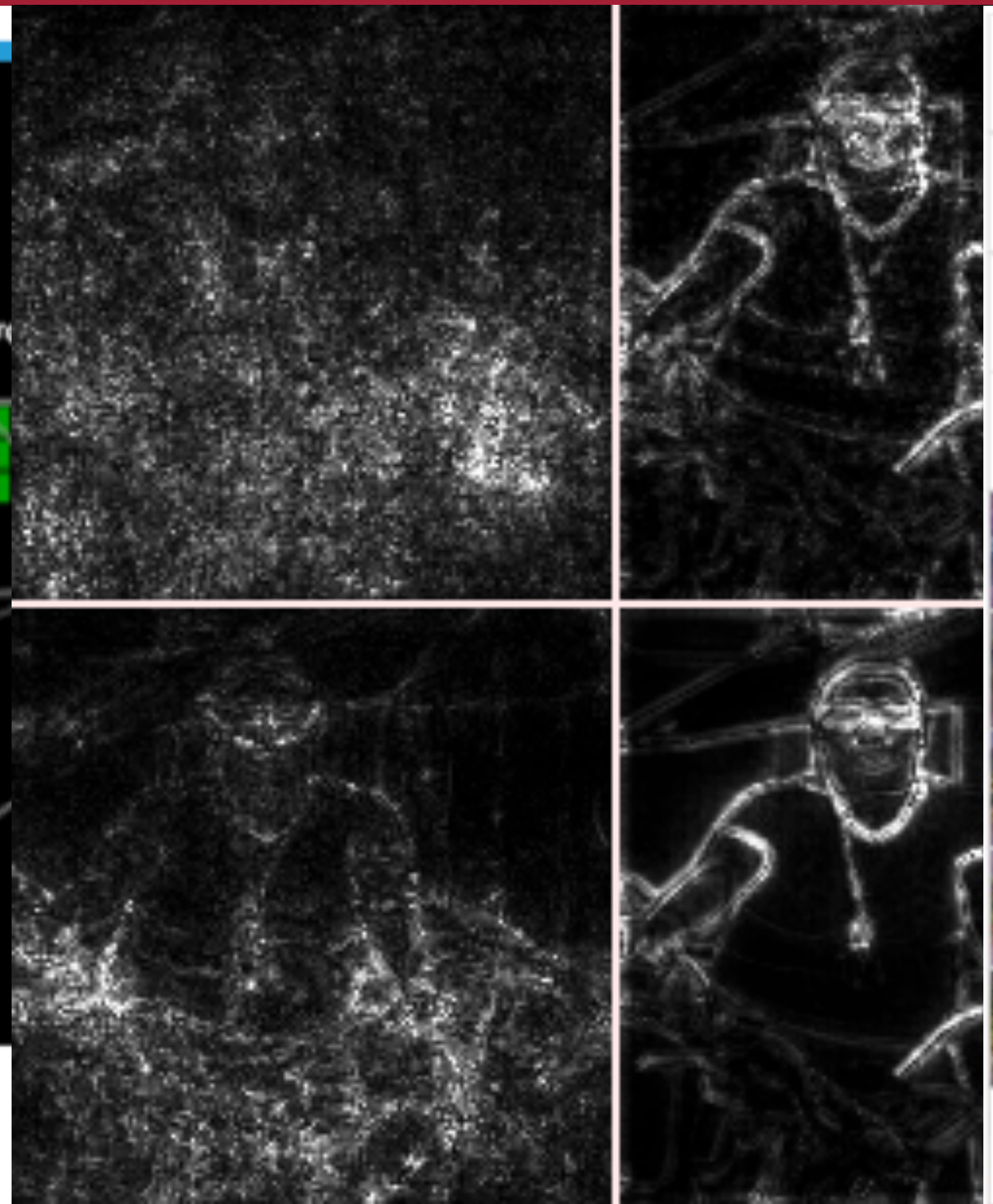
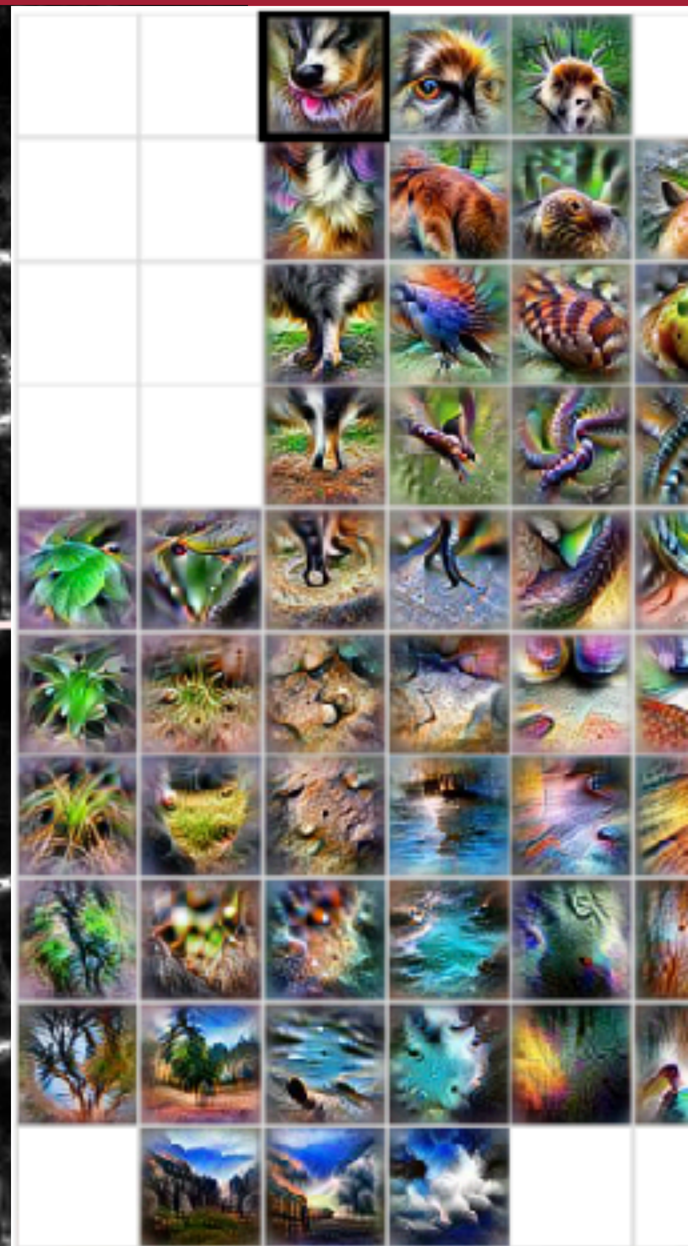
6.859: Interactive Data Visualization

Research Frontiers: ML Interpretability

Arvind Satyanarayan

Source: ImageNet

“Labrador Retriever”

“Golden Retriever”

Source: ImageNet

Source: ImageNet

“Labrador Retriever”

“Golden Retriever”

Label	Probability
Labrador retriever	69.2%
Golden retriever	11.6%
Tennis ball	2.6%

Why do we want interpretability?

Fairness: is the model biased? is it discriminating?

Causality: using models to infer properties about the natural world.

Reliability: how well does this model generalize, or transfer to a new domain?

Trust: how much confidence do I have in the model?

Transparency: how do I audit a model's decision-making?

[Lipton, 2017]


ICE - CLEAN CYCLING COPY
Start
Session Features
Browse...
New...
Cycling Keywords
Cycling Clubs and Organizations
Cycling Lessons
Bike Safety and Maintenance
Cycling Tours and Events
motorcycle
Inactive Features
Shopping and Rentals
shop, store, rental, rentals, produ...
FEATURES: 6 ITEMS: 511
Find Items
Sample
0.75
Find Diverse
Find False Positives
Find False Negatives
Submit
Found items around score of 0.75.
http://www.tillamookrodeo.com
Tillamook County Rodeo
Welcome to the 26th Annual 2012 Tillamook County Rodeo
ROBERT WARREN TRUCKING
SATURDAY AND SUNDAY JUNE 23 & 24, 2012
0.25
http://www.70wildmiles.org/
70 Wild Miles
A charity triathlon; 70 miles, 1 day, 3 challenges.
FERMANAGH
GLENCOE
BALA
0.75
Find Items
Featuring
Feature Suggestions
Review Labels
Standard Metrics
TEST
TRAIN
THRESHOLD: 0.0
1.0
41
223
38
209
Predicted
Positive
Negative
Actual
34
7
11
27
TEST
185
38
54
155
TRAIN
TEST AUC
0.76 (+0.00)
TRAIN AUC
0.88 (+0.00)
Test
Train
Precision-Recall Curves
Standard Metrics
ModelTracker [0.00, 1.00]
TEST AUC = 0.76 (+0.00) TRAIN AUC = 0.88 (+0.00)
3/6/2015 1:18:43 PM Error Friday, March 6, 2015 1:18:43 PM: ServerPerformanceService: ModelException: Unknown | (0): ServerPerformanceService: Failed to get aggregate counters
Stanford
University
1:18 PM
3/6/2015


ICE - CLEAN CYCLING COPY
Start
Session Features
Browse...
New...
Cycling Keywords
Cycling Clubs and Organizations
Cycling Lessons
Bike Safety and Maintenance
Cycling Tours and Events
motorcycle
Inactive Features
Shopping and Rentals
shop, store, rental, rentals, produ...
Review Labels
Score
All Data
Errors
Refresh View
Save Changes
0.94
0.92
/ 256
Find Items
Featuring
Feature Suggestions
Review Labels
Standard Metrics
TEST TRAIN
THRESHOLD: 0.0
1.0
41 223
38 209
Predicted
Positive Negative
Actual
34 7
11 27
TEST
185 38
54 155
TRAIN
TEST AUC
0.76 (+0.00)
TRAIN AUC
0.88 (+0.00)
Test Train
Precision-Recall Curves
Standard Metrics
ModelTracker [0.00, 1.00]
FEATURES: 6 ITEMS: 511
TEST AUC = 0.76 (+0.00) TRAIN AUC = 0.88 (+0.00)
3/6/2015 1:20:08 PM Error Friday, March 6, 2015 1:20:08 PM: ServerPerformanceService: ModelException: Unknown | (0): ServerPerformanceService: Failed to get aggregate counters
1:20 PM
3/6/2015
Stanford
University

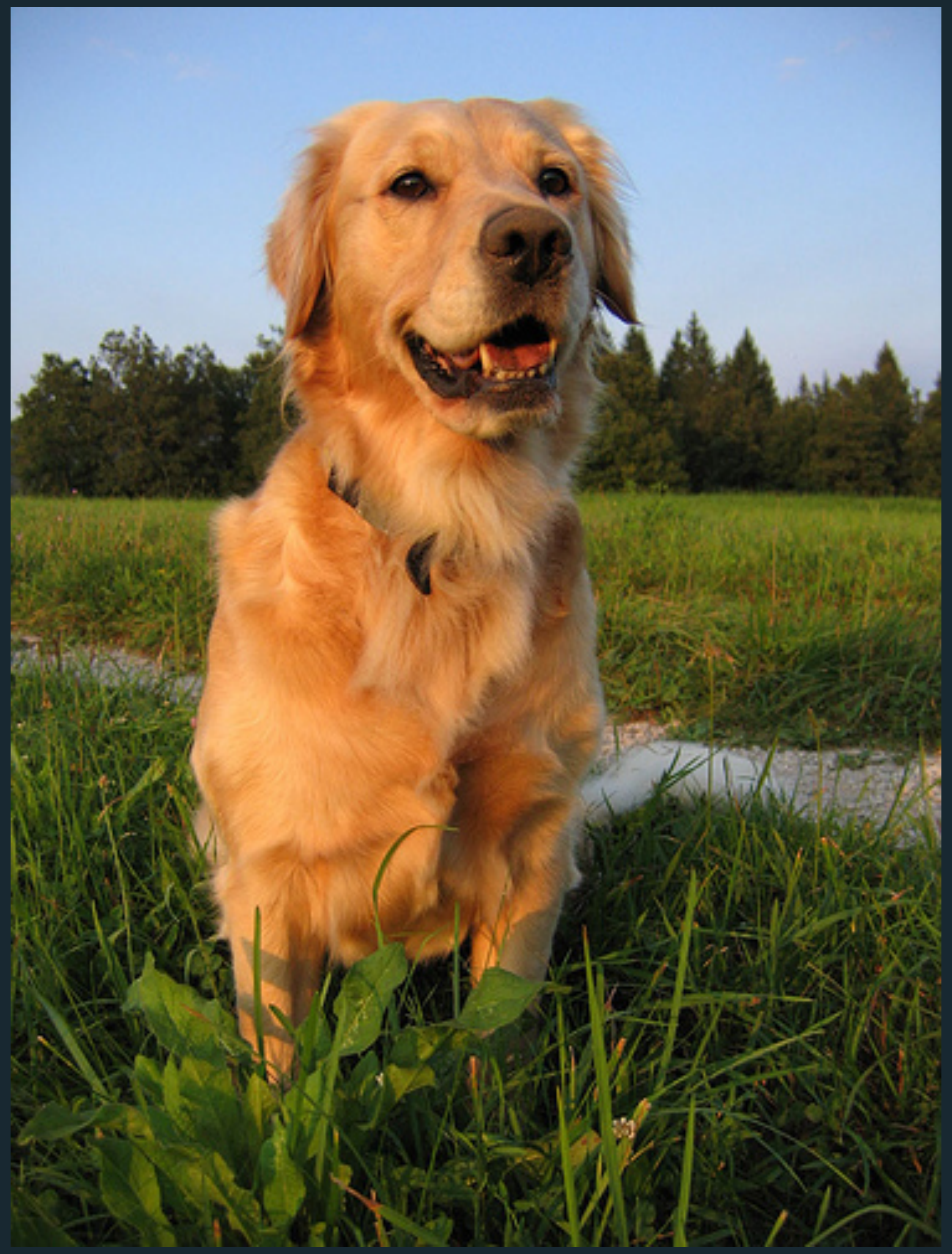

Stanford
University

ModelTracker

[Amershi et al., CHI 2015]

Source: ImageNet

“Labrador Retriever”



“Golden Retriever”

Label	Probability
Labrador retriever	69.2%
Golden retriever	11.6%
Tennis ball	2.6%

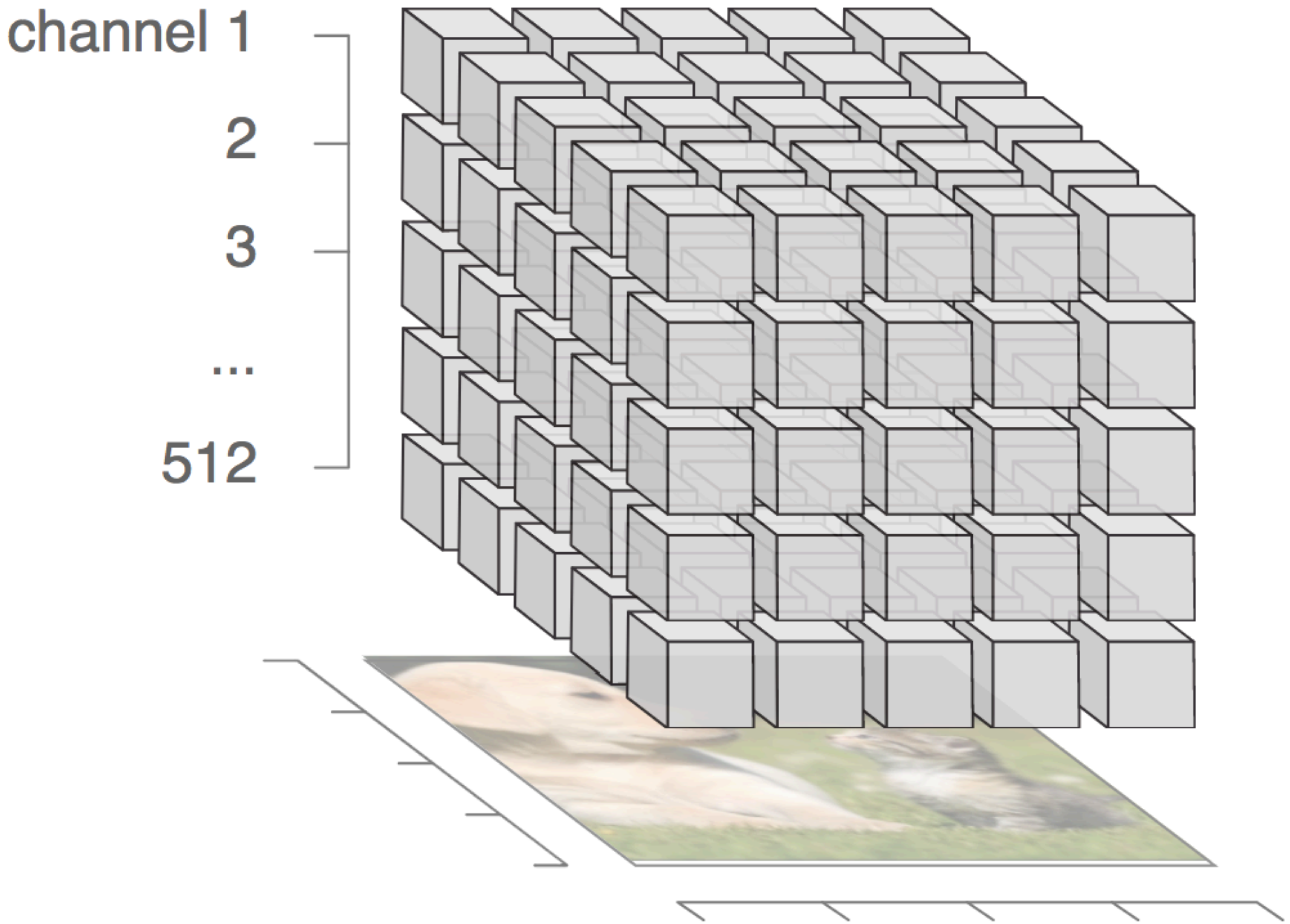
Hidden Layers

“Labrador Retriever”

“Golden Retriever”

Label	Probability
Labrador retriever	69.2%
Golden retriever	11.6%
Tennis ball	2.6%

Individual Neurons



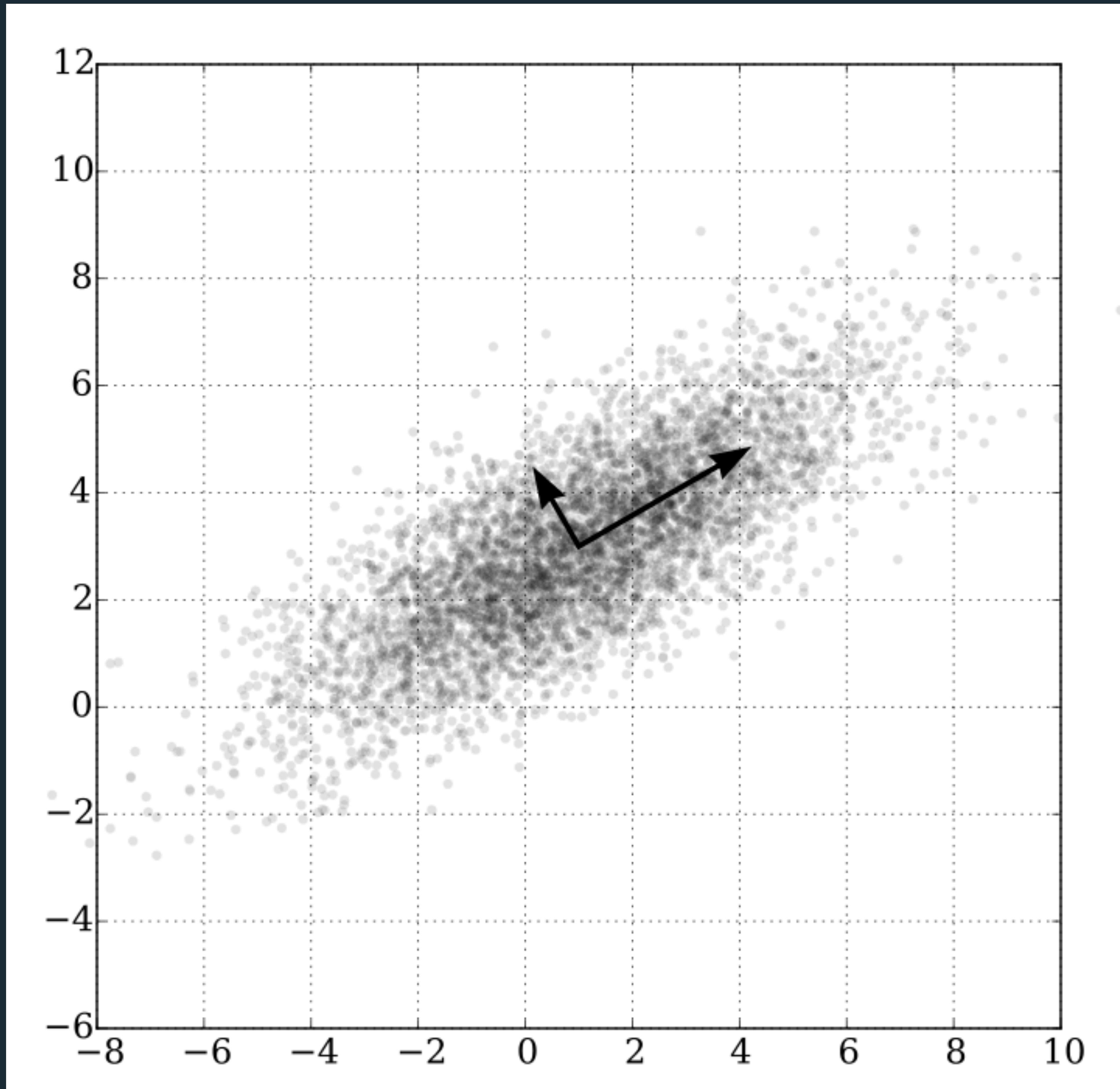
Dimensionality Reduction

Project nD data to 2D or 3D. Interpret/sanity check learned representations.

But **subject to their own interpretation issues**.

Different techniques make different trade-offs:

PCA (Principal Component Analysis): roughly fit a p-dimensional ellipsoid to the data, order axes by amount of data variance they explain. Preserves global structure.



Dimensionality Reduction

Project nD data to 2D or 3D. Interpret/sanity check learned representations.

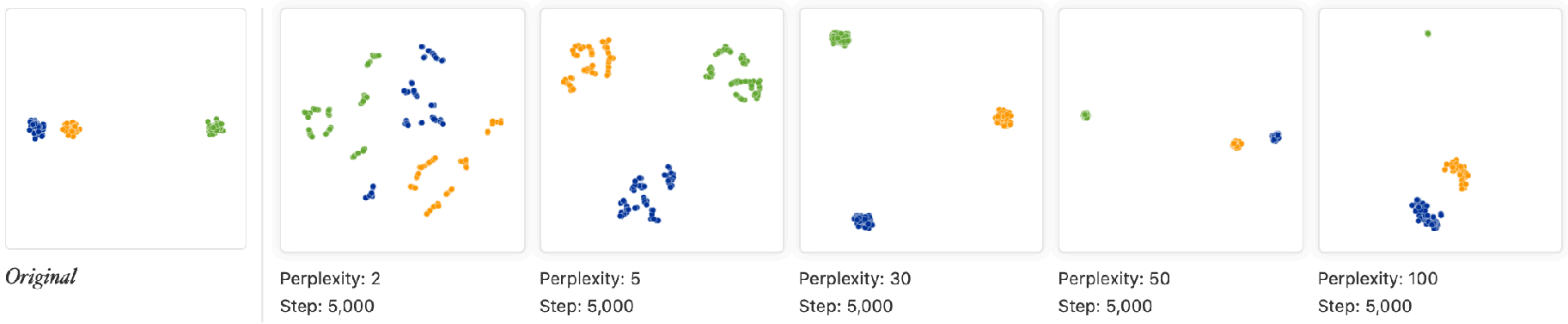
But **subject to their own interpretation issues**.

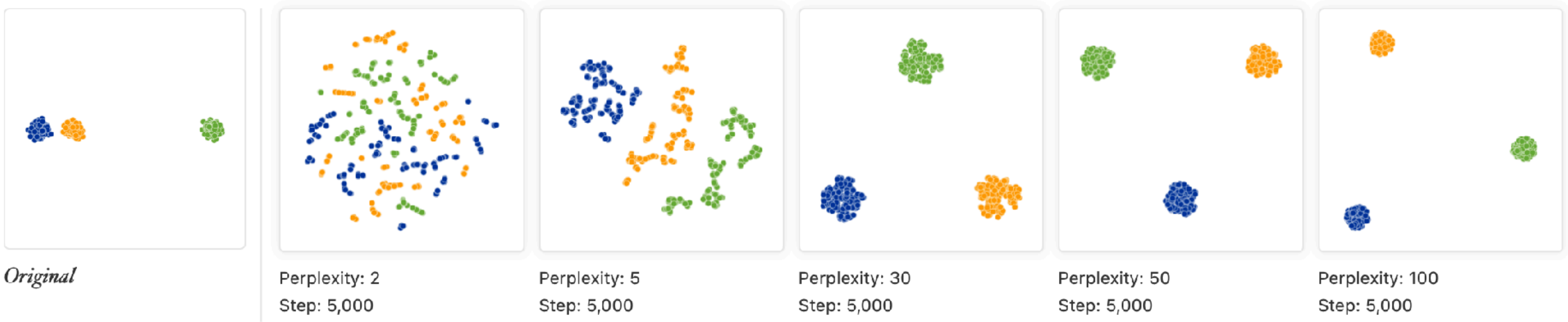
Different techniques make different trade-offs:

- **PCA (Principal Component Analysis)**: roughly fit a p-dimensional ellipsoid to the data, order axes by amount of data variance they explain. Preserves global structure.

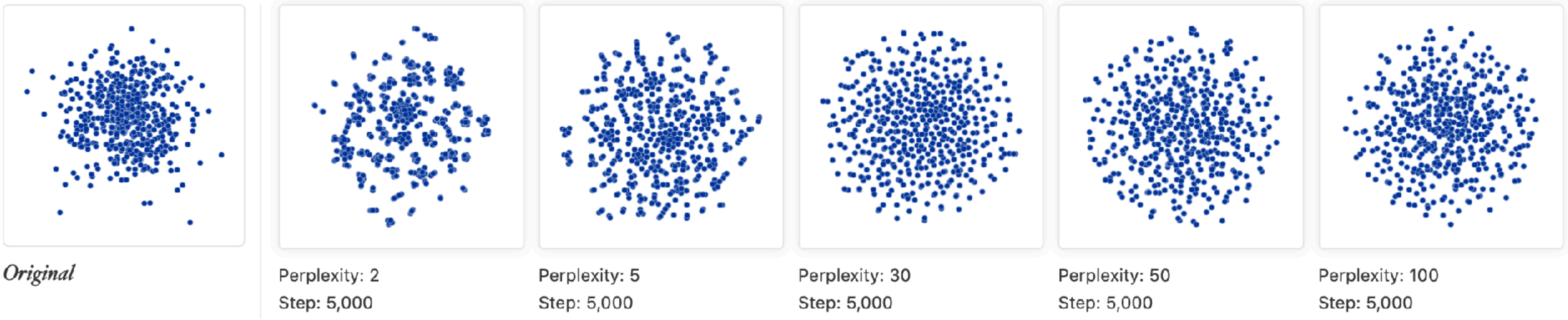
- **t-SNE (t-Dist. Stochastic Neighbor Embedding)**: probabilistic distribution that adapts and performs different transformations on different regions.

[Wattenberg et al., Distill 2016]





[Wattenberg et al., Distill 2016]



[Wattenberg et al., Distill 2016]

Dimensionality Reduction

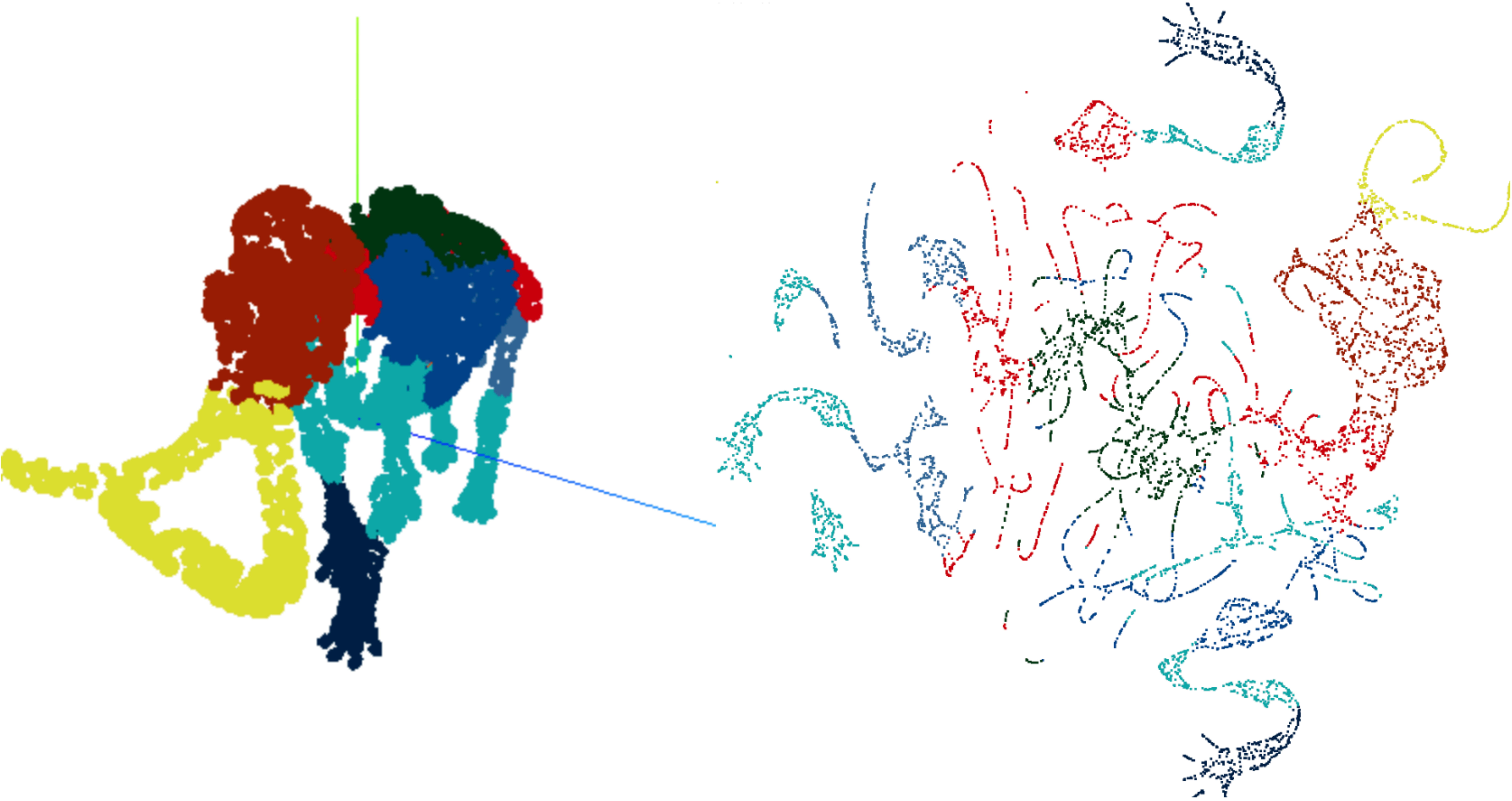
Project nD data to 2D or 3D. Interpret/sanity check learned representations.

But **subject to their own interpretation issues**.

Different techniques make different trade-offs:

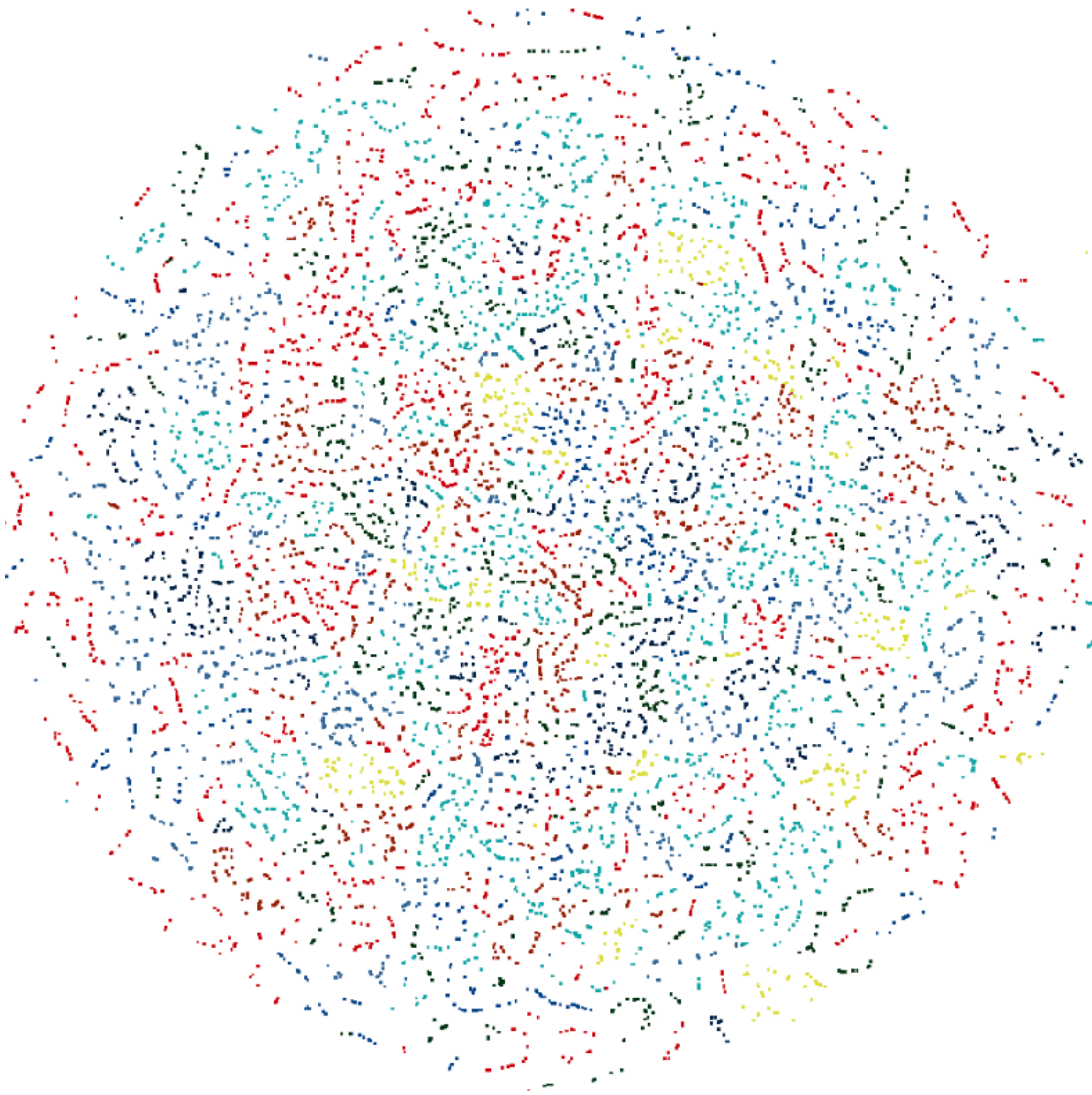
> **PCA (Principal Component Analysis)**: roughly fit a p-dimensional ellipsoid to the data, order axes by amount of data variance they explain. Preserves global structure.
>
> **t-SNE (t-Dist. Stochastic Neighbor Embedding)**: probabilistic distribution that adapts and performs different transformations on different regions.
>
> **UMAP (Uniform Manifold Approx. & Projection)**: Identify local regions, stitch them together. Tries to balance local/global trade-off.

Original 3D Data **2D UMAP Projection**



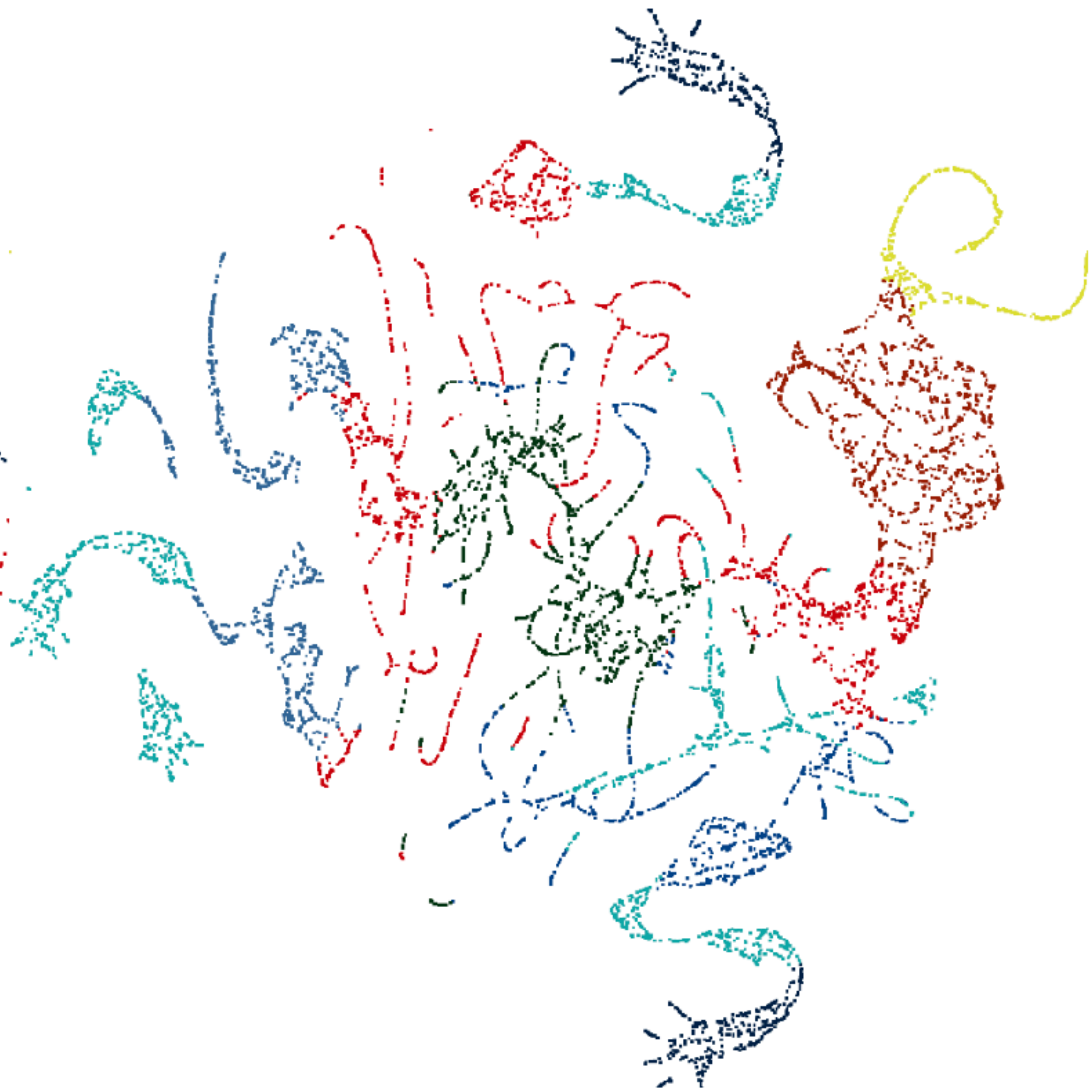
[Coenen & Pearce, 2019]

2D t-SNE projection



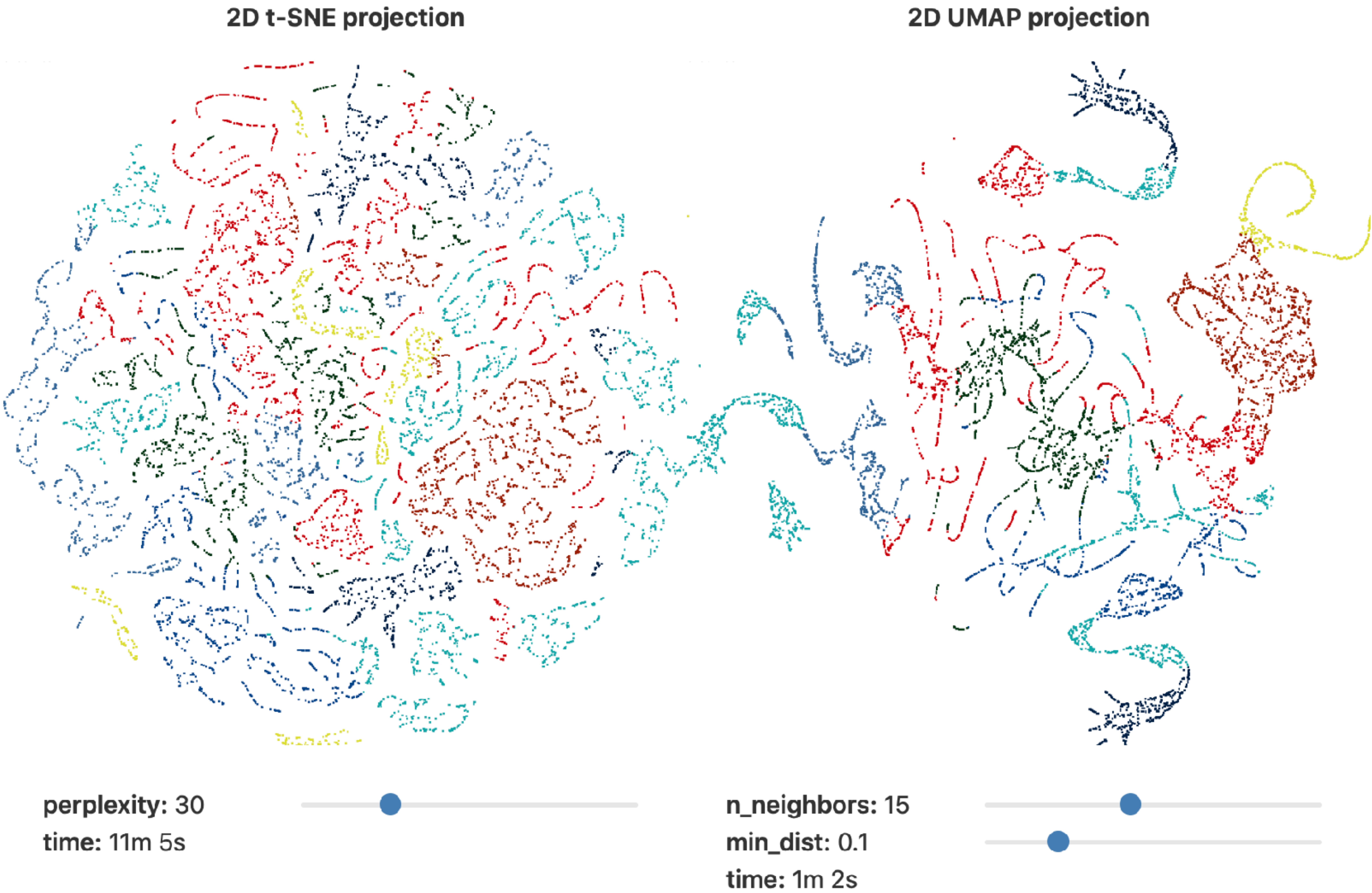
perplexity: 5
time: 9m 18s

2D UMAP projection



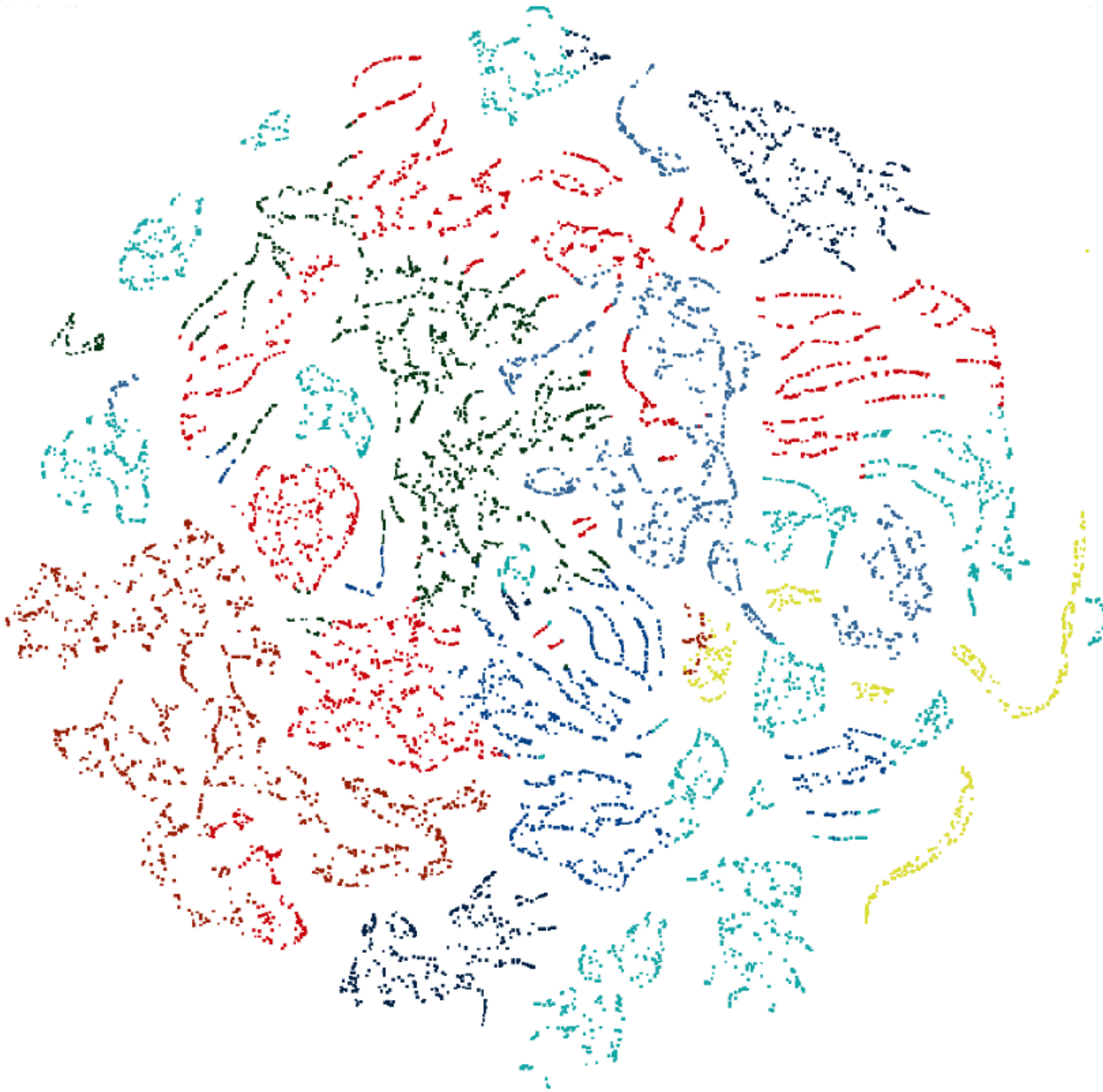
n_neighbors: 15
min_dist: 0.1
time: 1m 2s

[Coenen & Pearce, 2019]



[Coenen & Pearce, 2019]

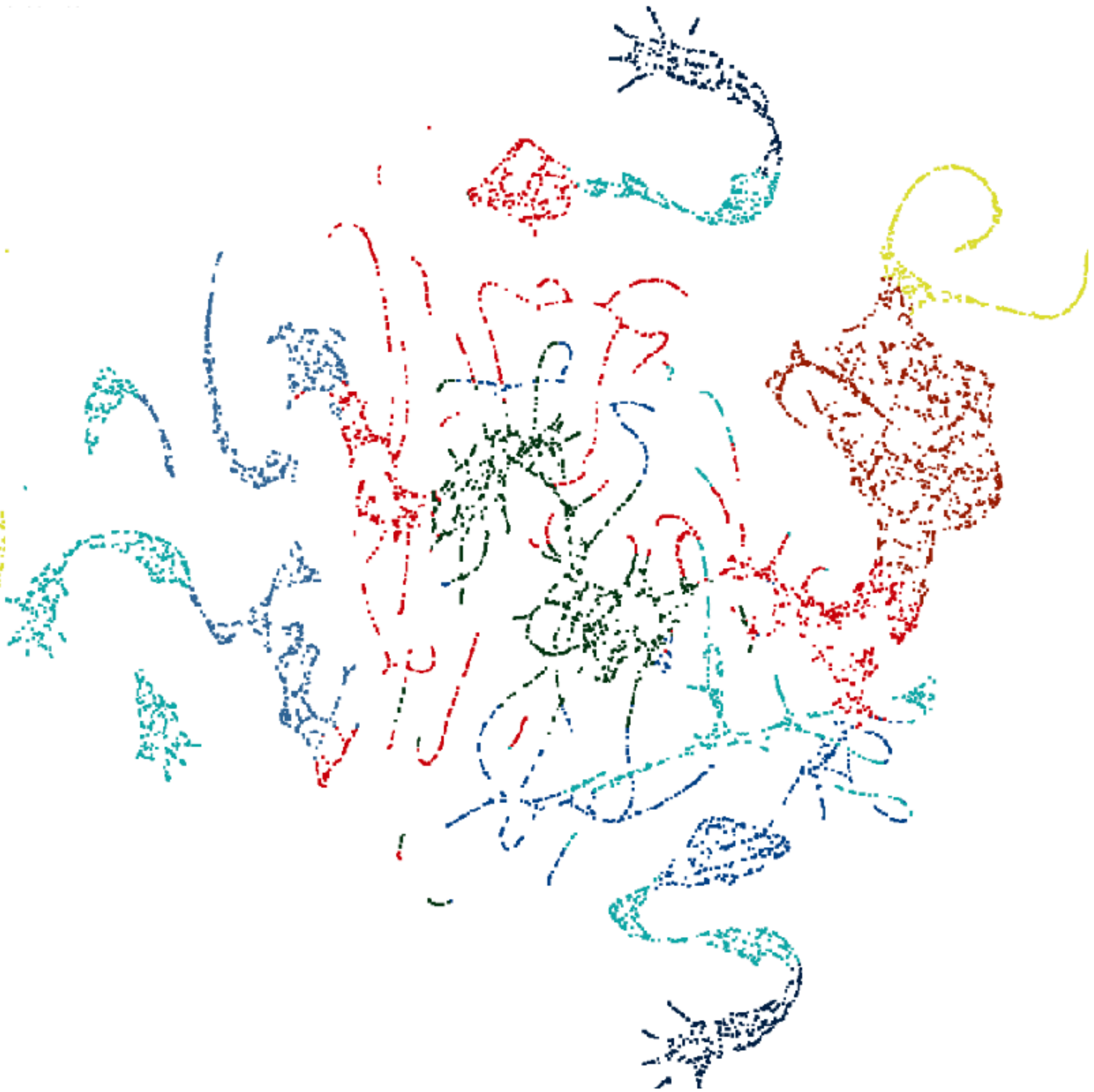
2D t-SNE projection



perplexity: 50

time: 12m 42s

2D UMAP projection

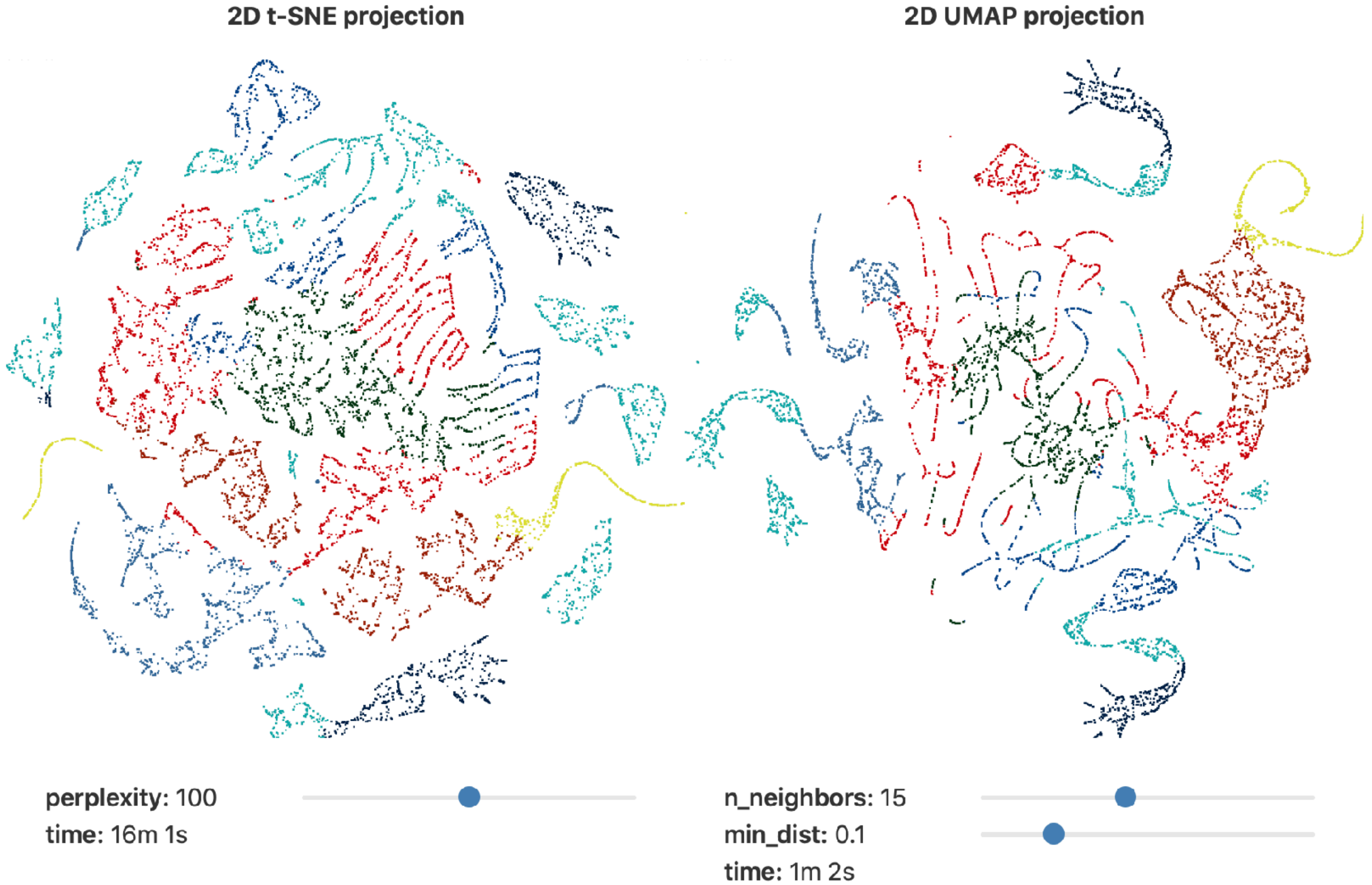


n_neighbors: 15

min_dist: 0.1

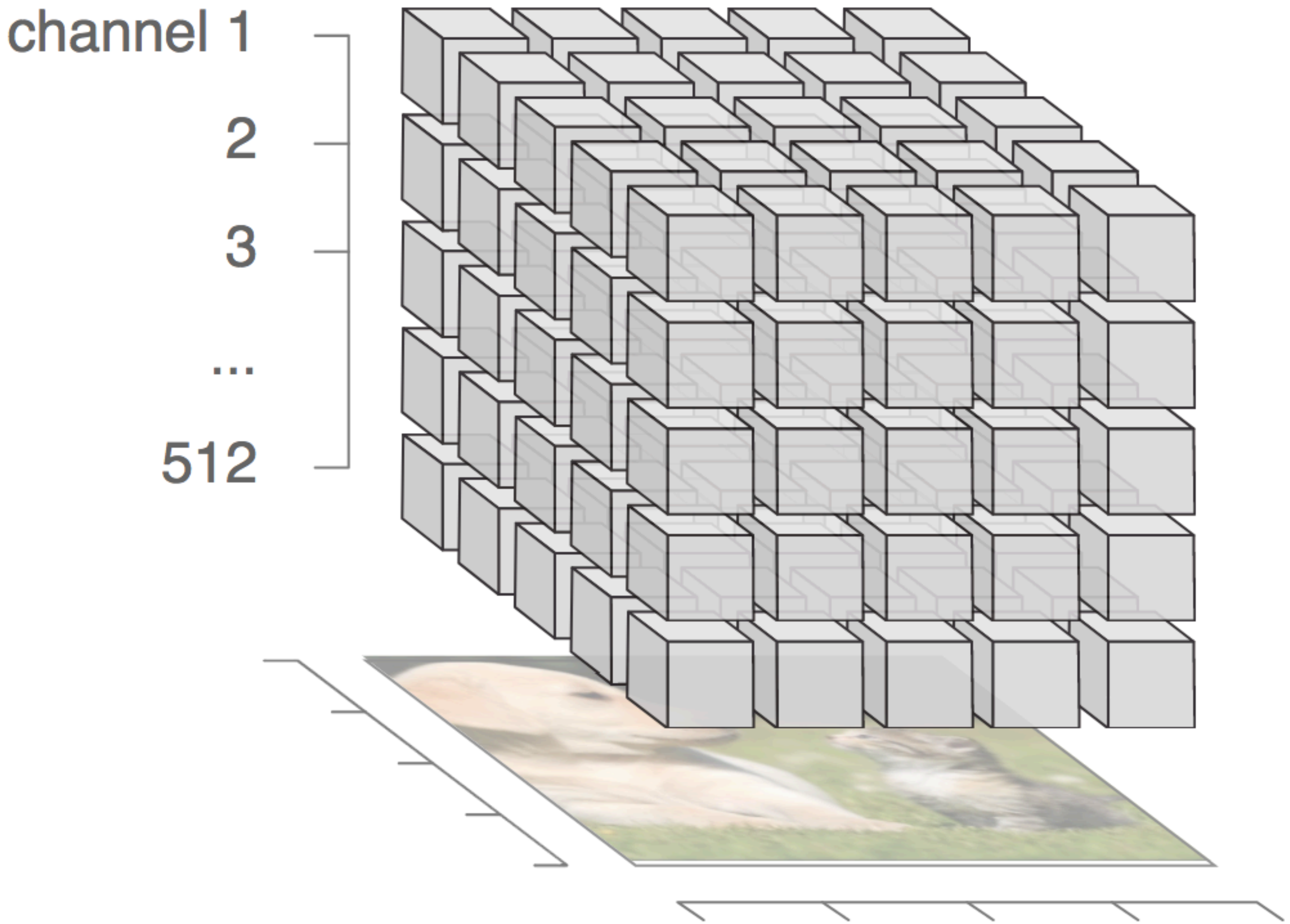
time: 1m 2s

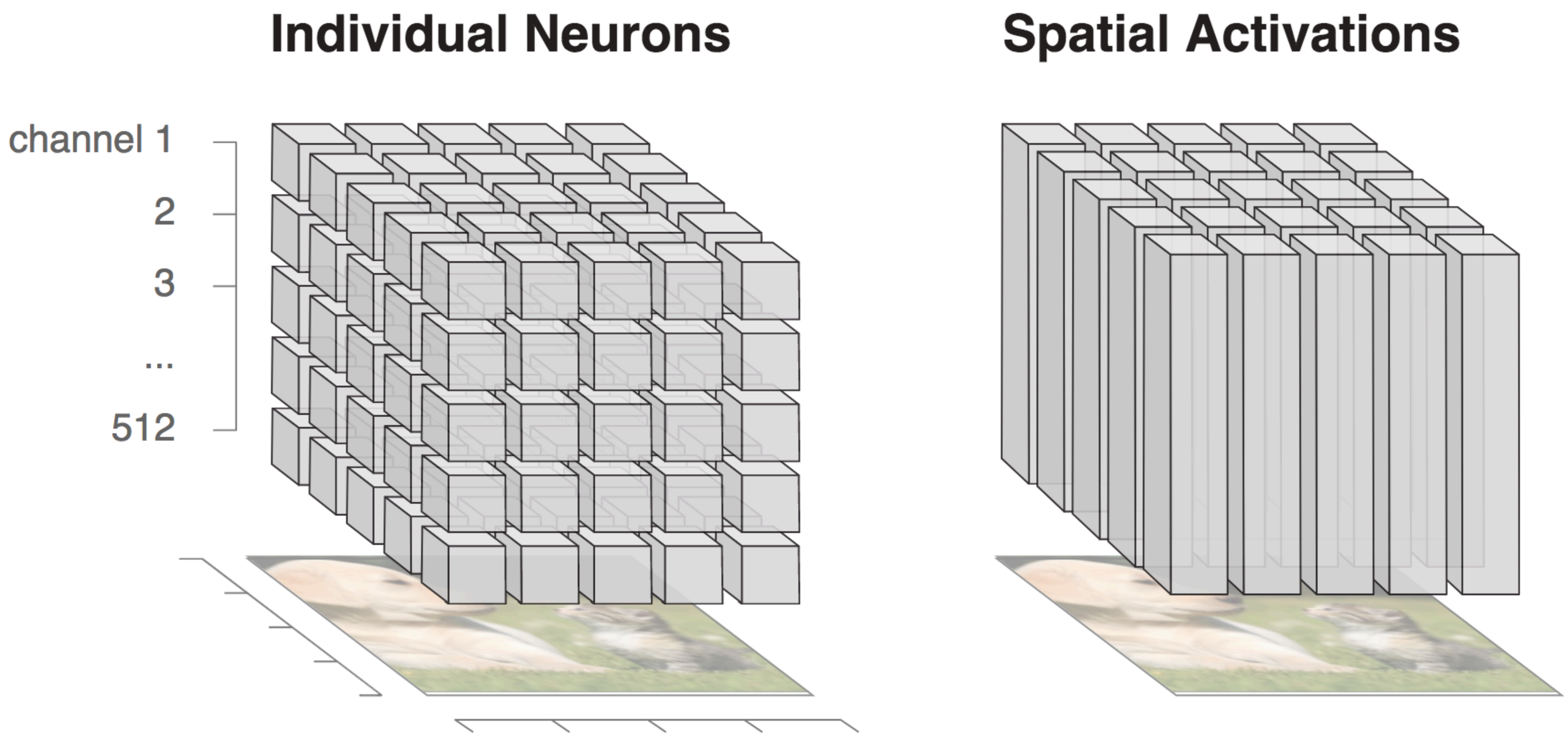
[Coenen & Pearce, 2019]



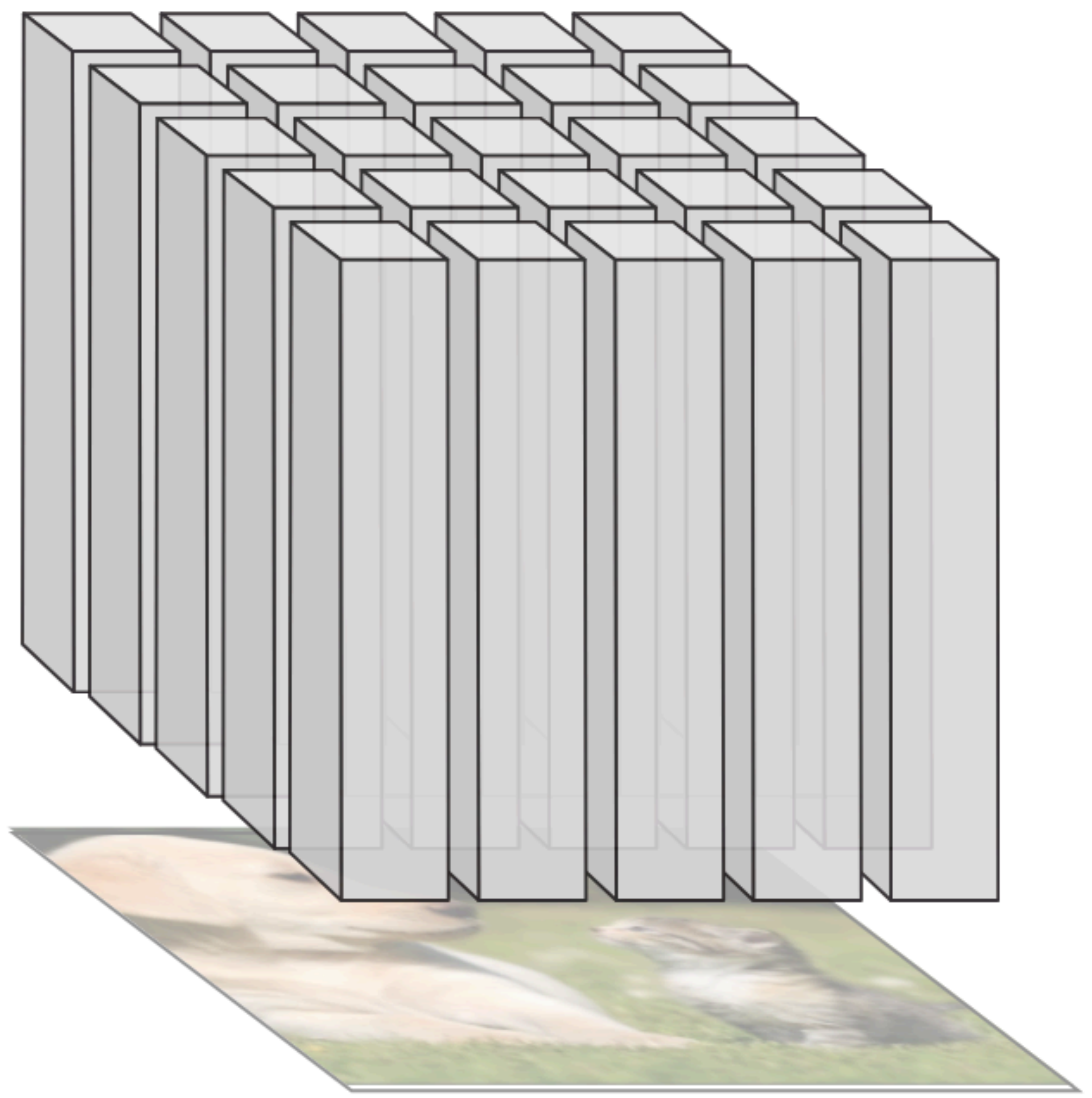
[Coenen & Pearce, 2019]

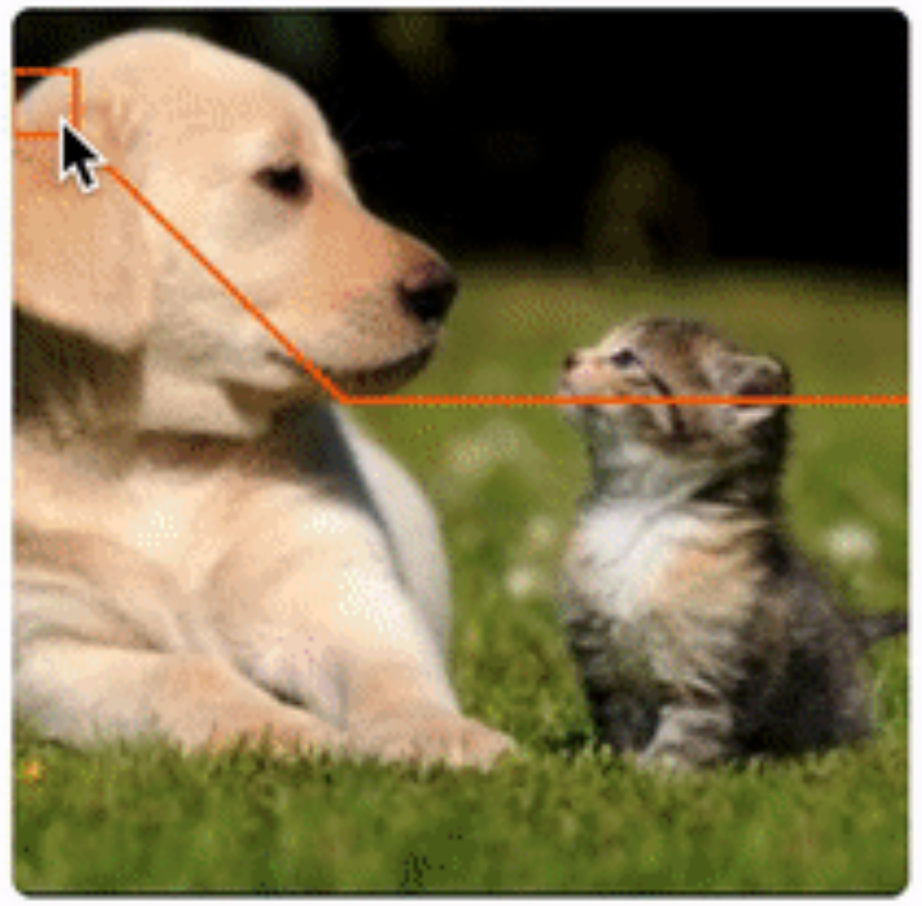
Individual Neurons




Individual Neurons
Spatial Activations
channel 1
2
3
...
512

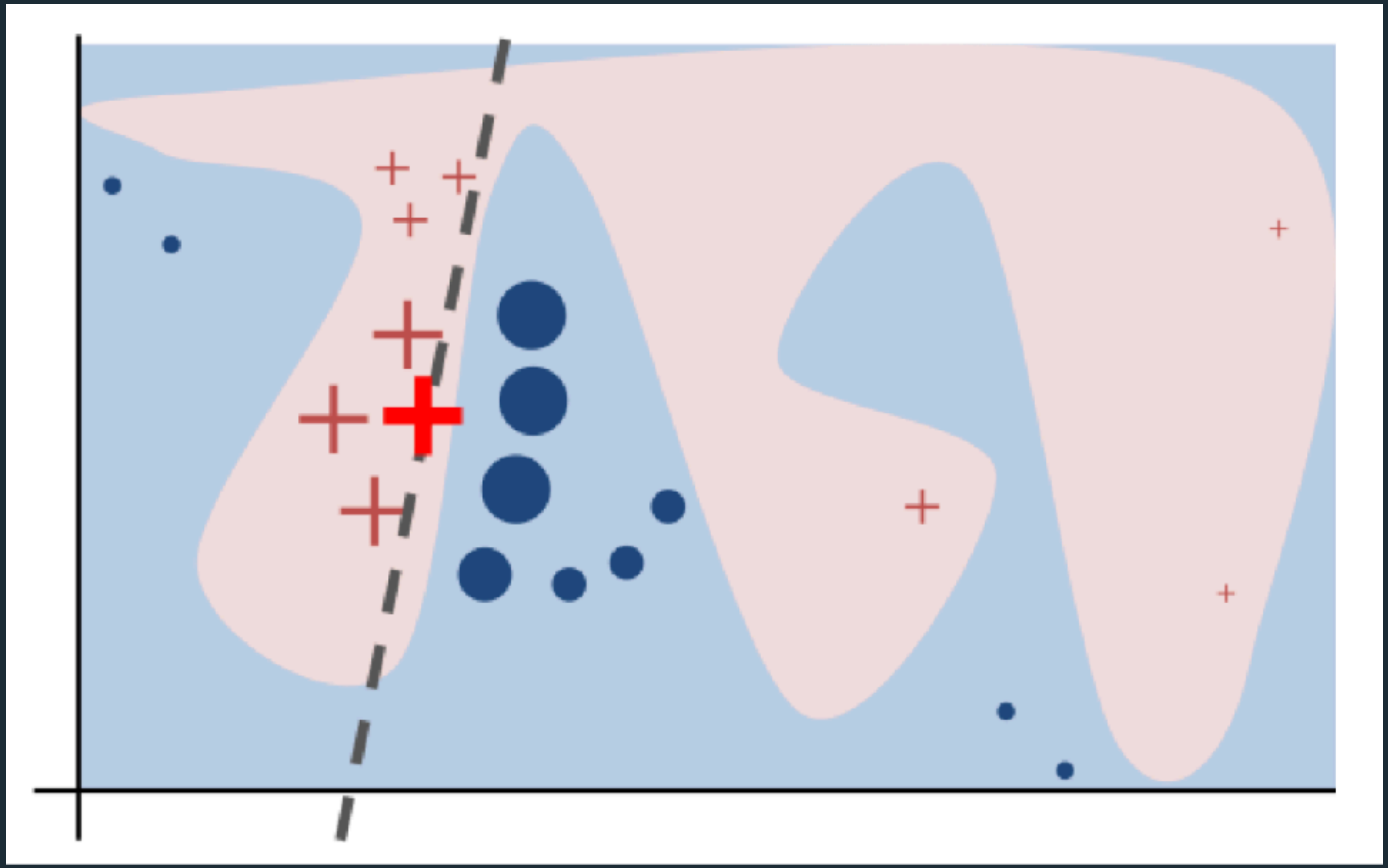
Spatial Activations



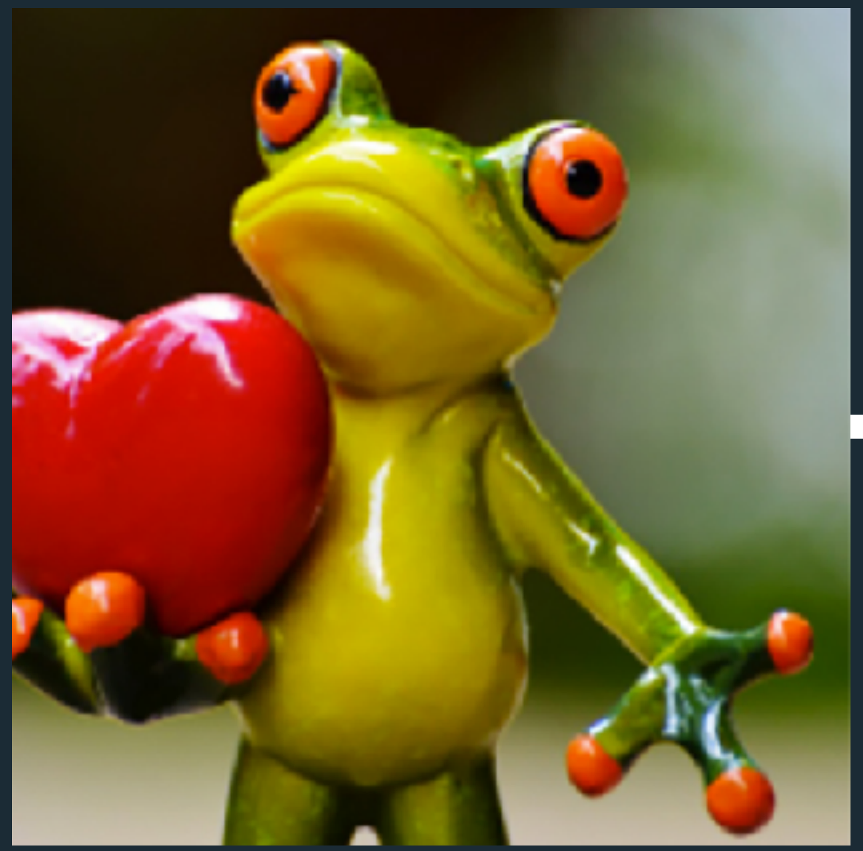


$a_{1,0}$ = [0, 0, 0, 0, 49.6, 0, 43.6, 30.2, 119.8, 62.7, 0, 51...

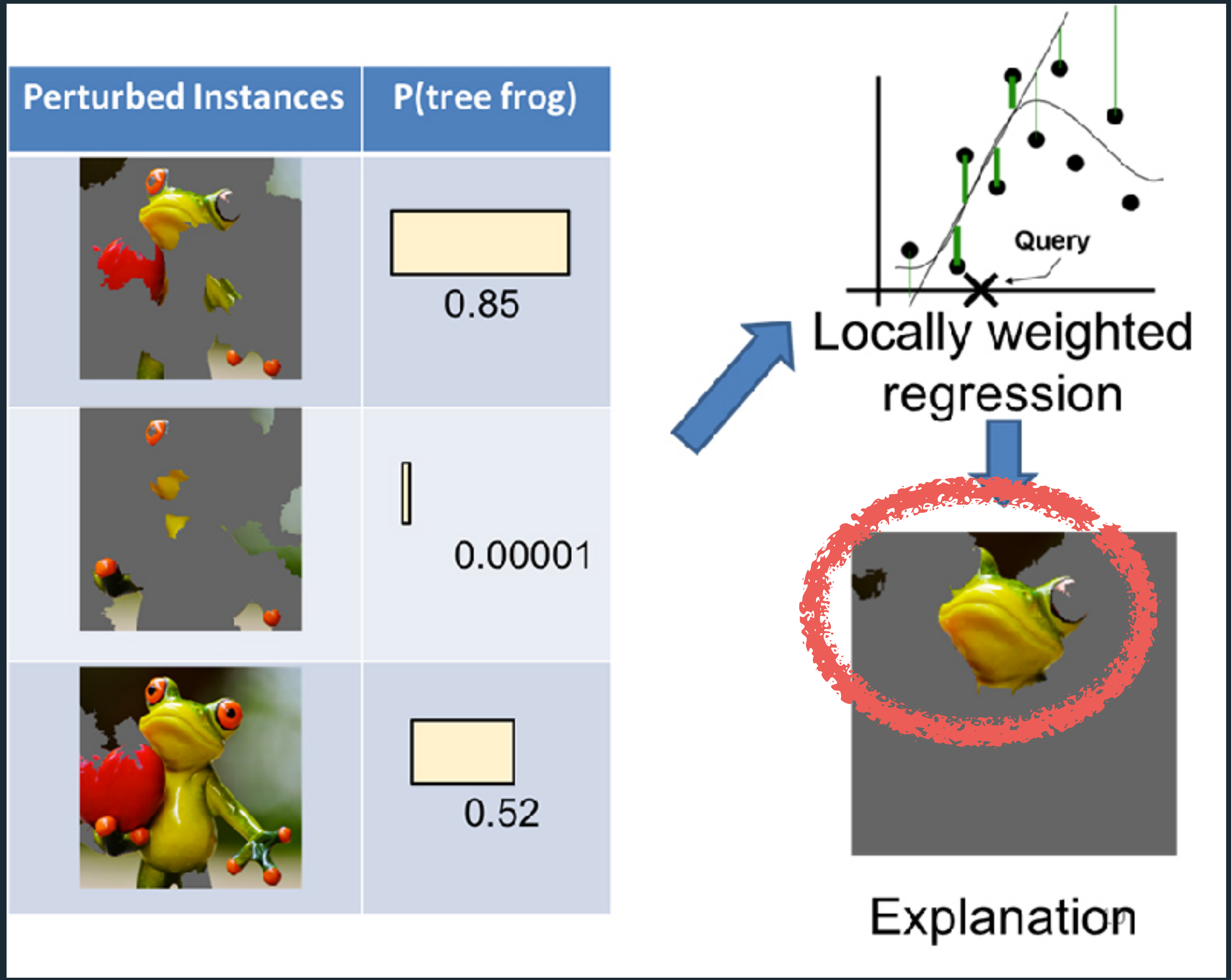
Local
Interpretable
Model-Agnostic
Explanations



[Ribeiro et al., KDD 2016]

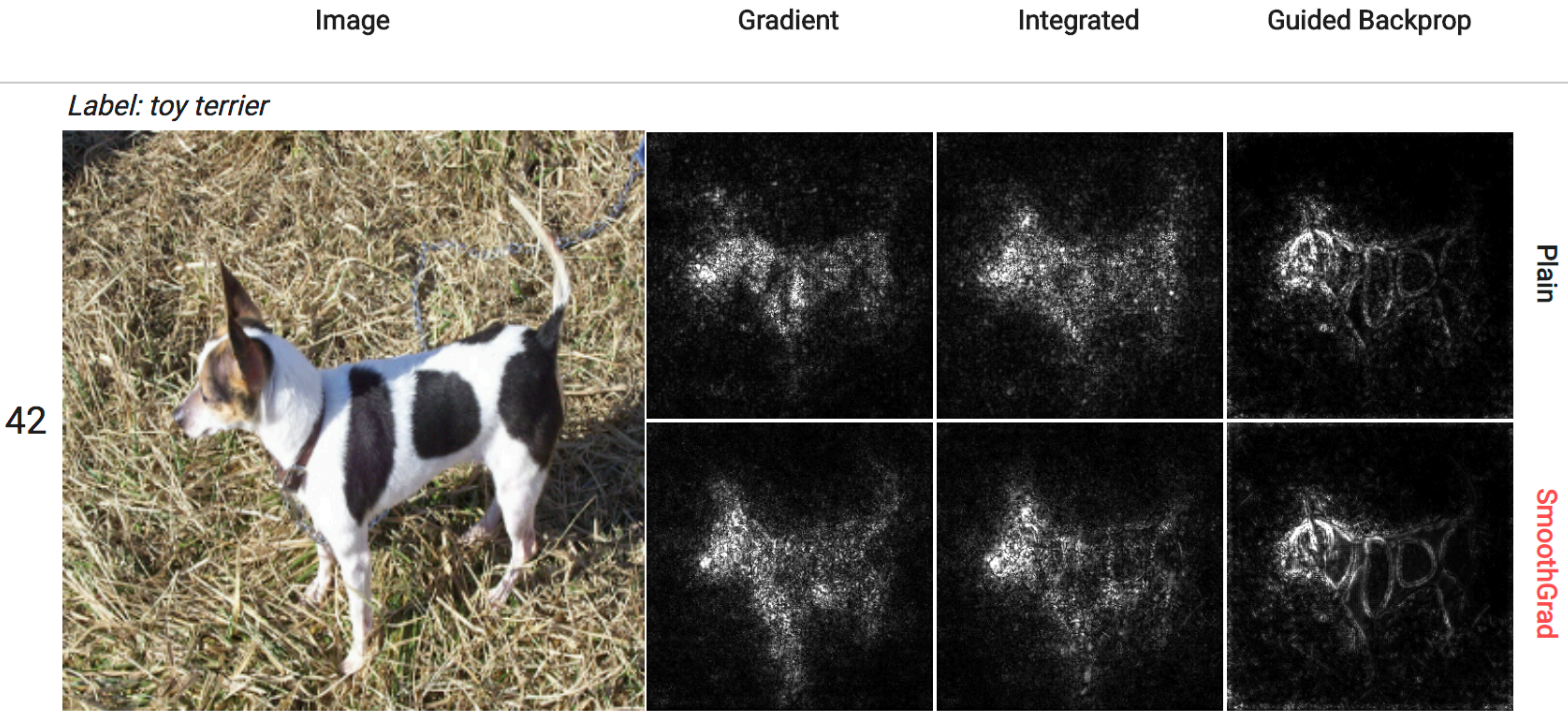


Local **I**nterpretable **M**odel-Agnostic **E**xplanations



Regions sufficient for "tree frog" classification.

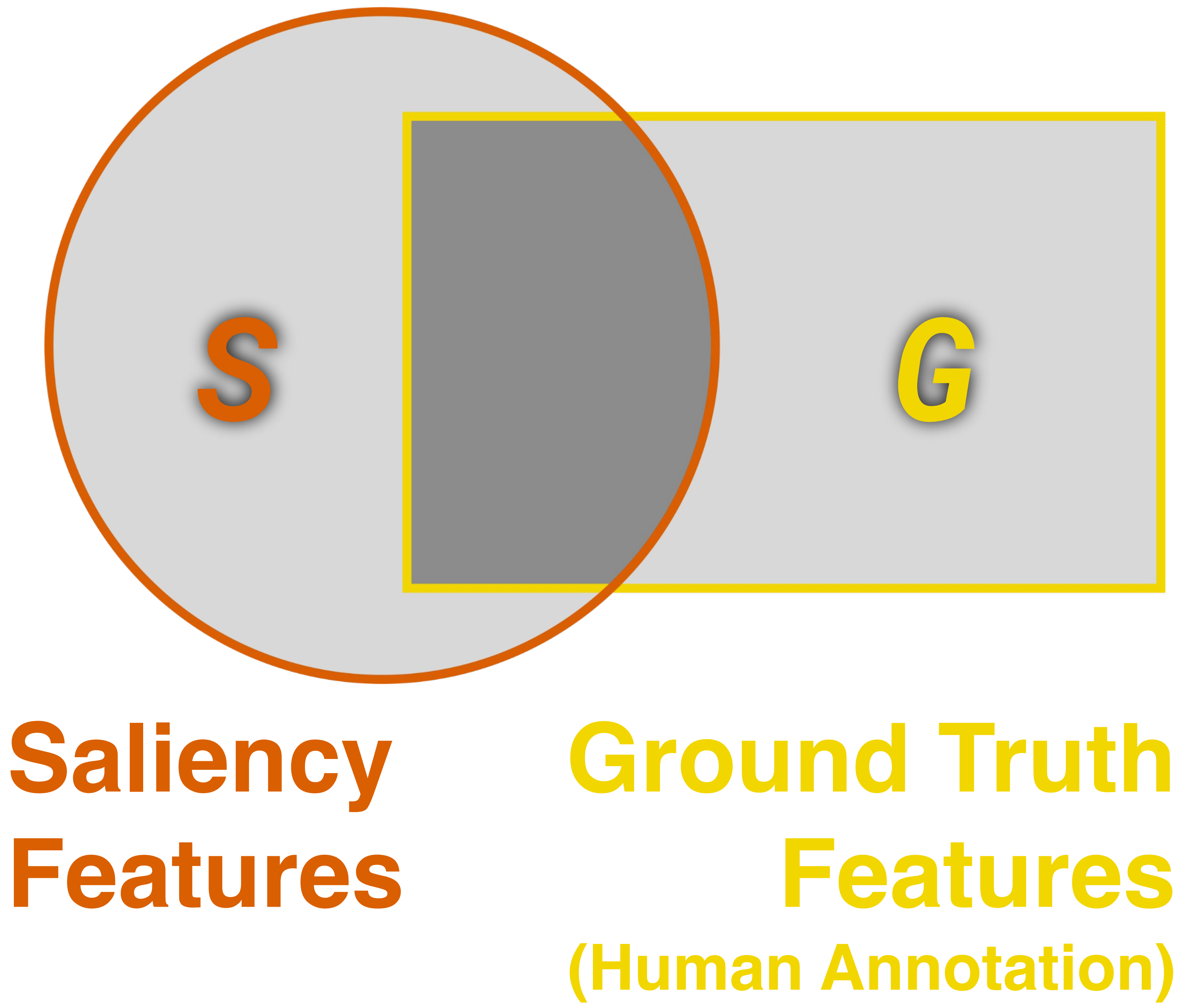
[Ribeiro et al., KDD 2016]



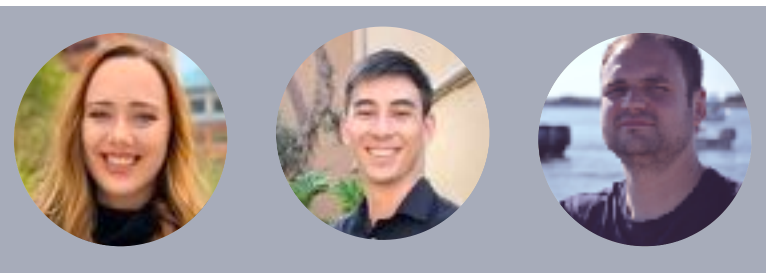
[SmoothGrad. Smilkov et al., 2017]

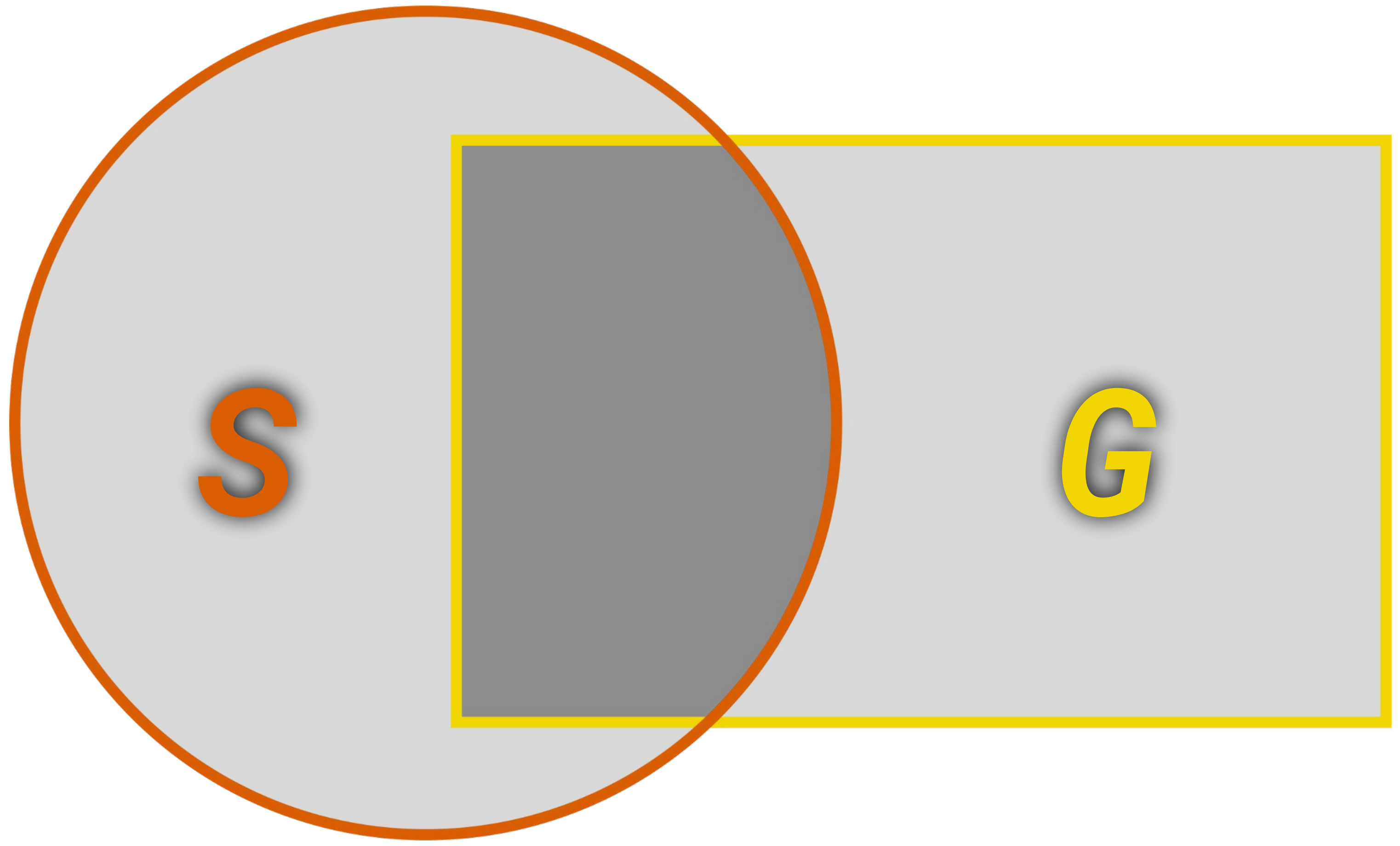
Shared Interest: Measuring Human-AI Alignment

with



Shared Interest: Measuring Human-AI Alignment

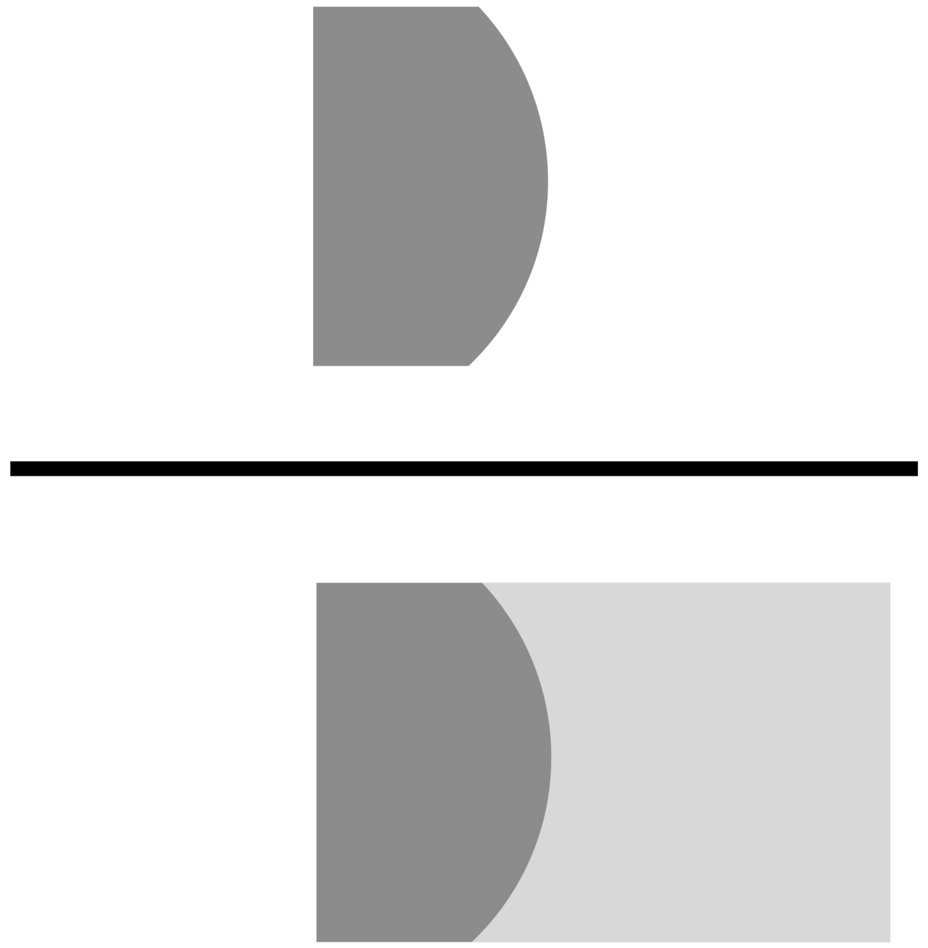
with 



Saliency Features

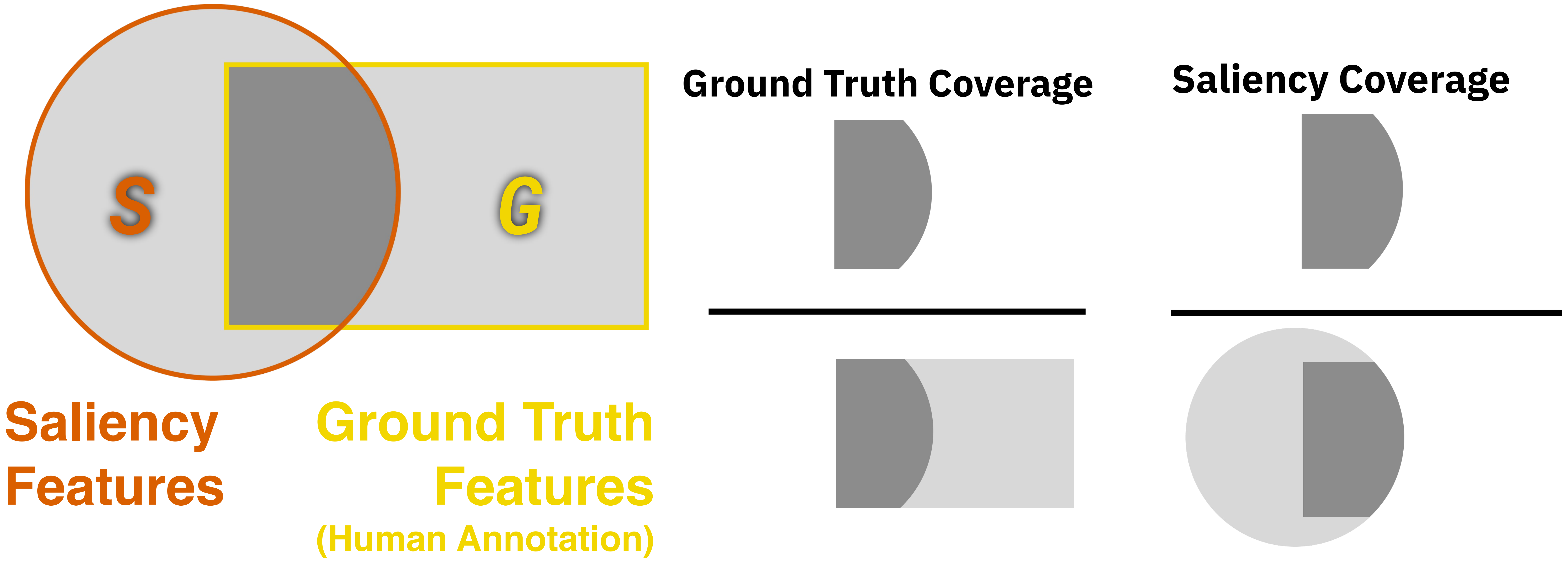
Ground Truth Features
(Human Annotation)

Ground Truth Coverage

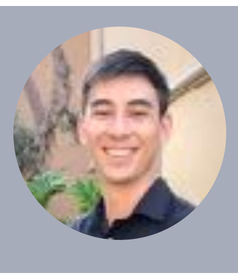


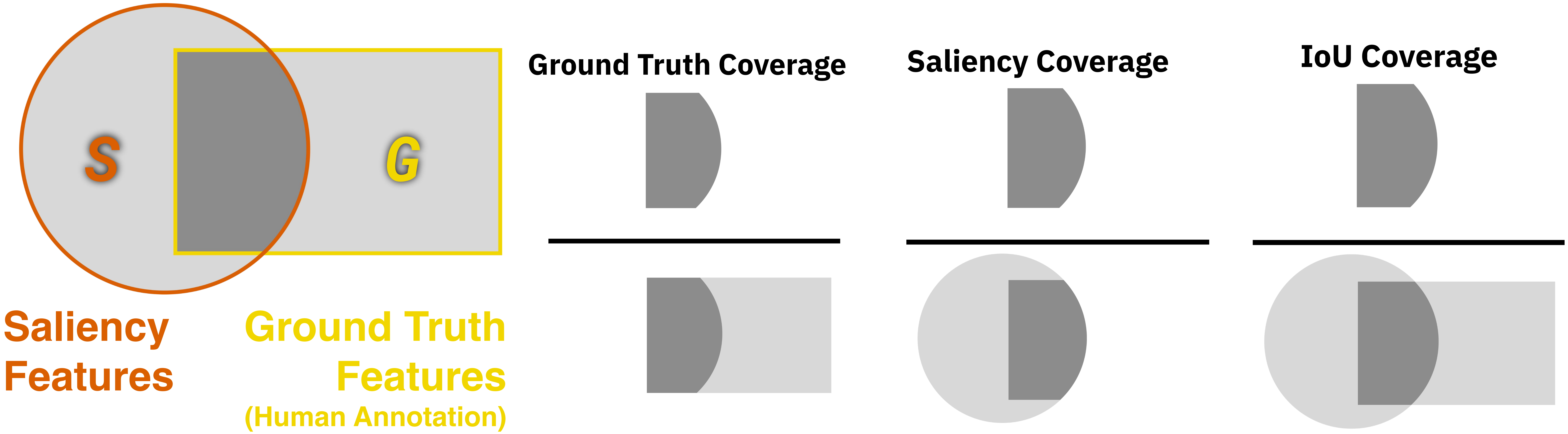
Shared Interest: Measuring Human-AI Alignment

with

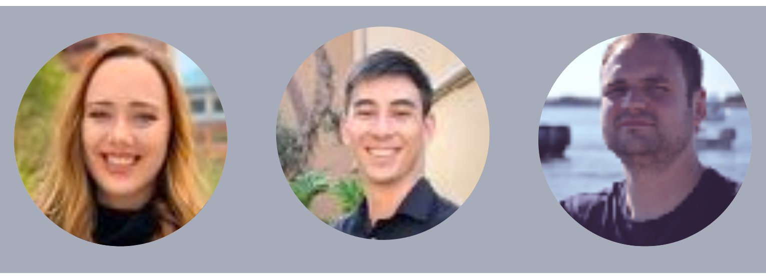


Shared Interest: Measuring Human-AI Alignment

with   



Shared Interest: Measuring Human-AI Alignment

with 

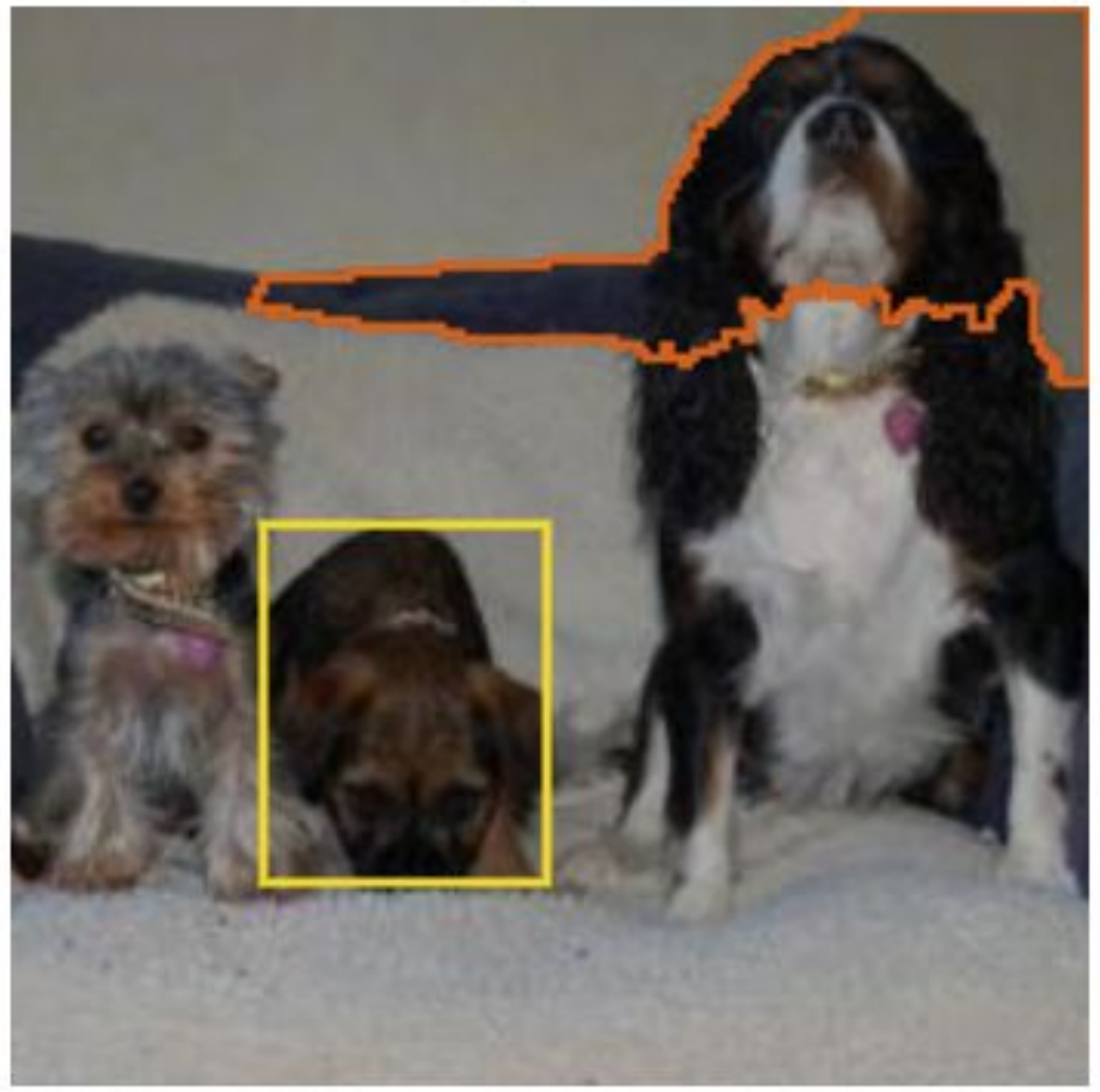
GROUND TRUTH COVERAGE

LOW SHARED INTEREST SCORE

Incorrect

IoU: 0.00 GTC: 0.00 SC: 0.00

border terrier | blenheim spaniel



Correct

IoU: 0.01 GTC: 0.02 SC: 0.04

mountain bike | mountain bike

HIGH SHARED INTEREST SCORE

Incorrect

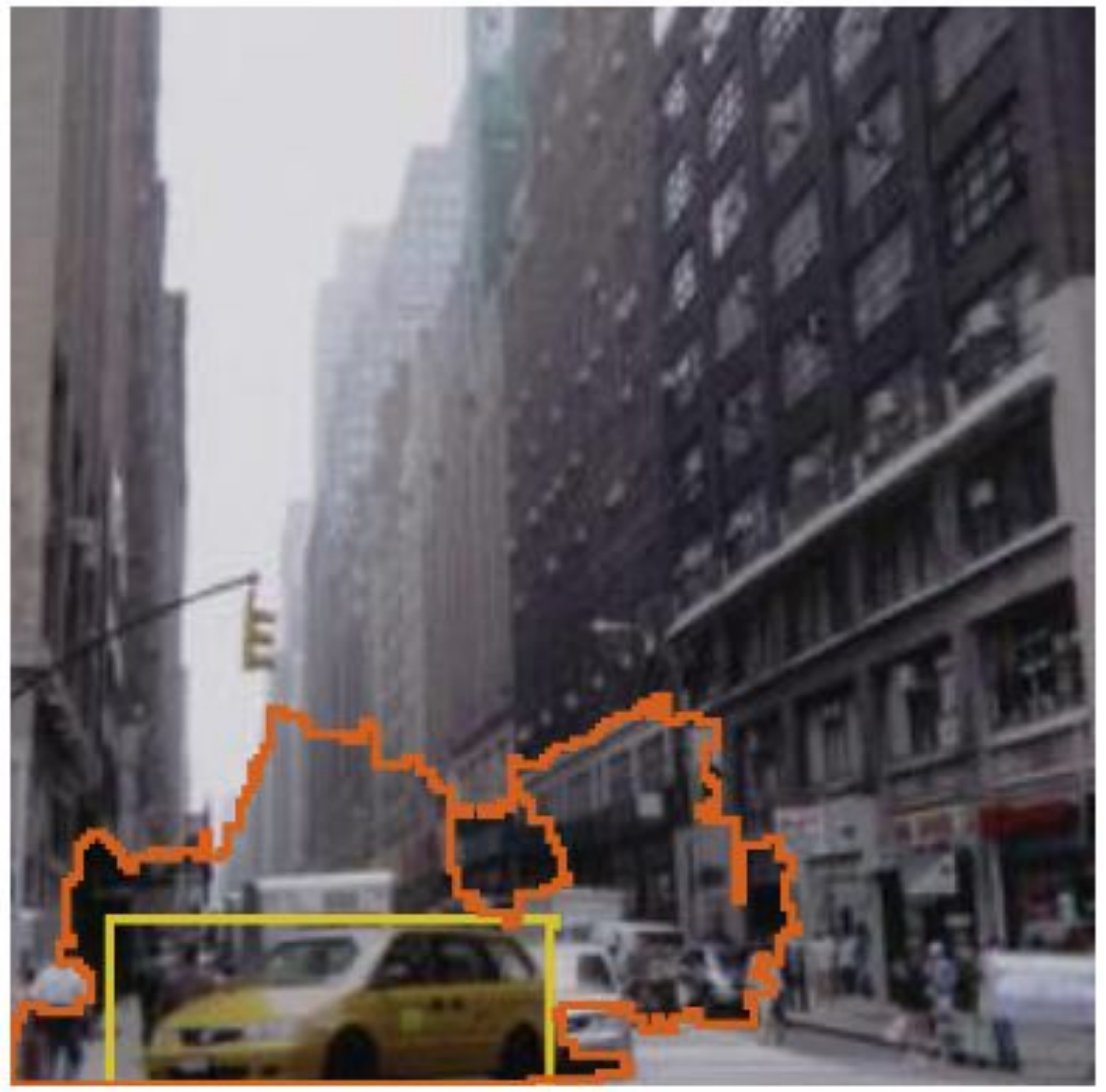
IoU: 0.21 GTC: 0.95 SC: 0.21

chihuahua | laptop

Correct

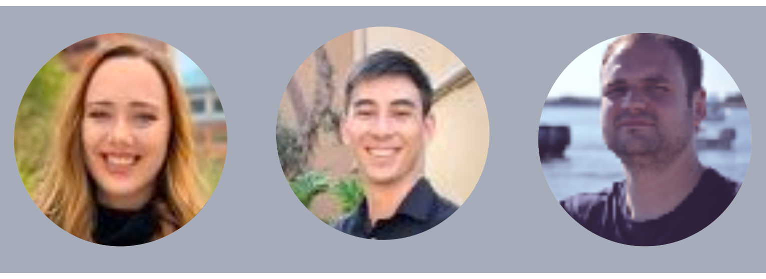
IoU: 0.38 GTC: 1.00 SC: 0.38

cab | cab



Shared Interest: Measuring Human-AI Alignment

with



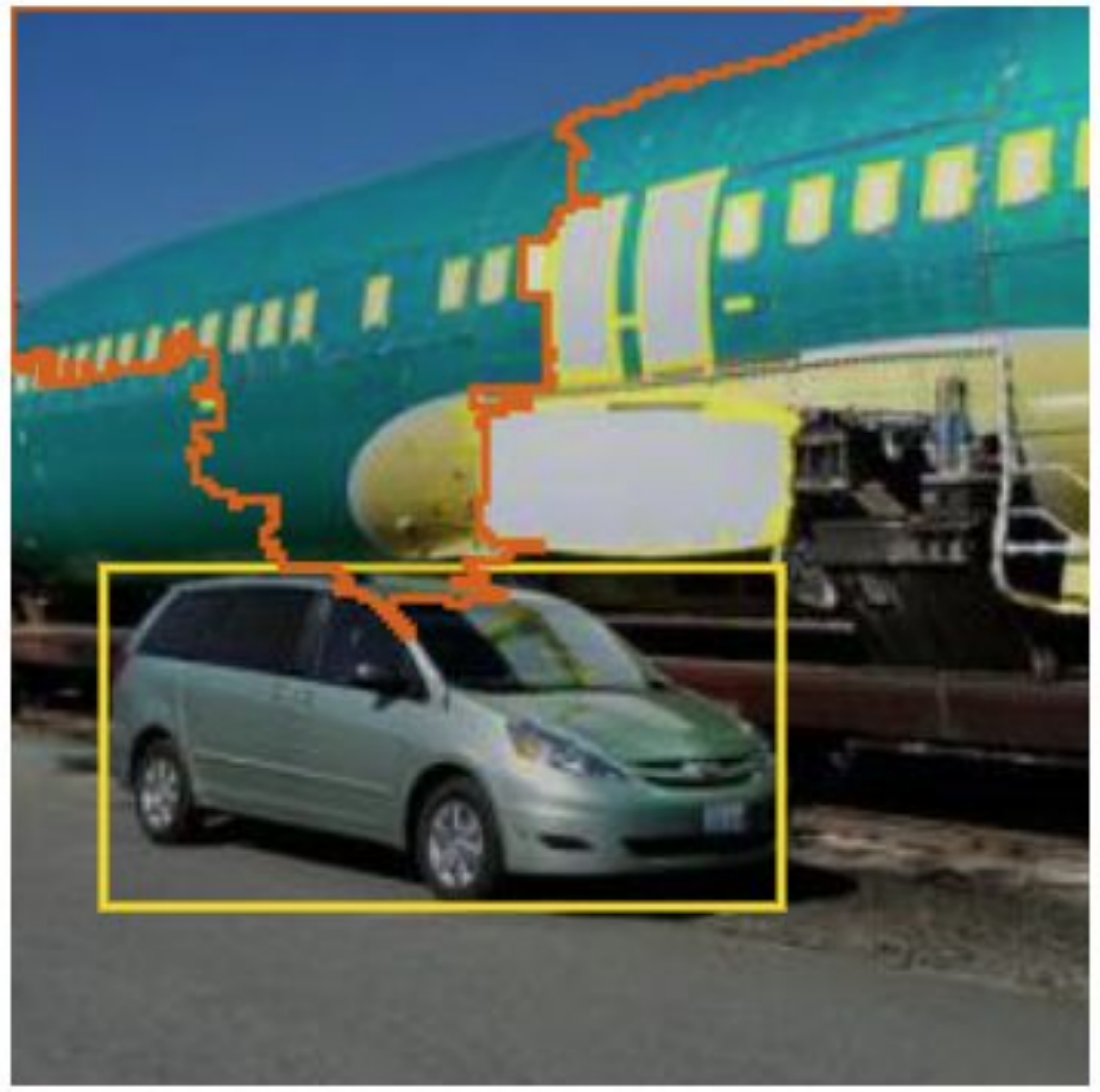
SALIENCY COVERAGE

LOW SHARED INTEREST SCORE

Incorrect

IoU: 0.01 | GTC: 0.02 | SC: 0.02

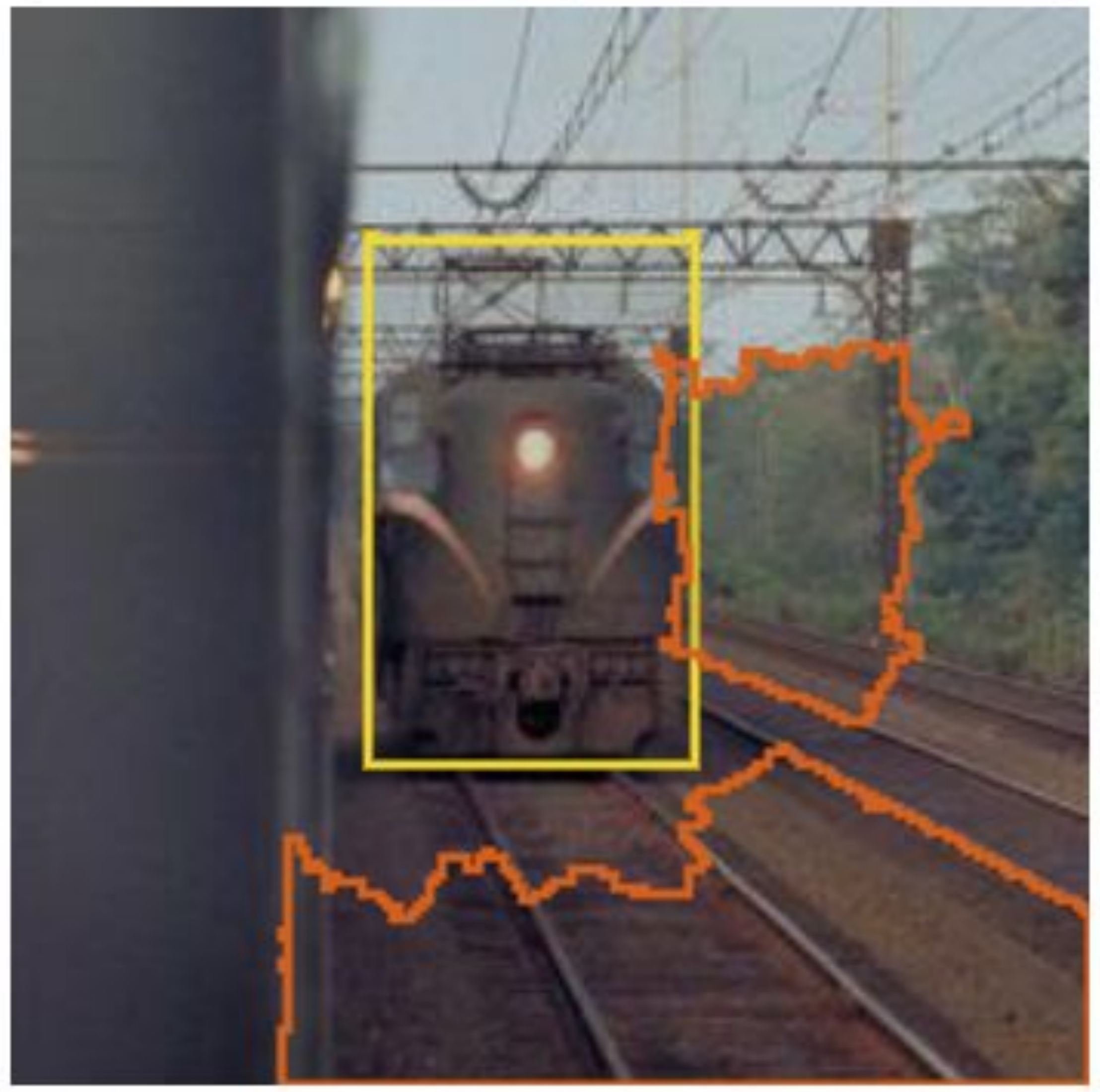
minivan | airliner



Correct

IoU: 0.02 | GTC: 0.04 | SC: 0.02

electric loco... | electric loco...



HIGH SHARED INTEREST SCORE

Incorrect

IoU: 0.14 | GTC: 0.14 | SC: 1.00

english setter | dalmatian

Correct

IoU: 0.31 | GTC: 0.31 | SC: 1.00

maltese dog | maltese dog

Shared Interest: Measuring Human-AI Alignment

with   

IOU COVERAGE

LOW SHARED INTEREST SCORE

Incorrect

IoU: 0.00 GTC: 0.00 SC: 0.00

samoyed arctic fox

Correct

IoU: 0.04 GTC: 0.05 SC: 0.17

horse cart horse cart

HIGH SHARED INTEREST SCORE

Incorrect

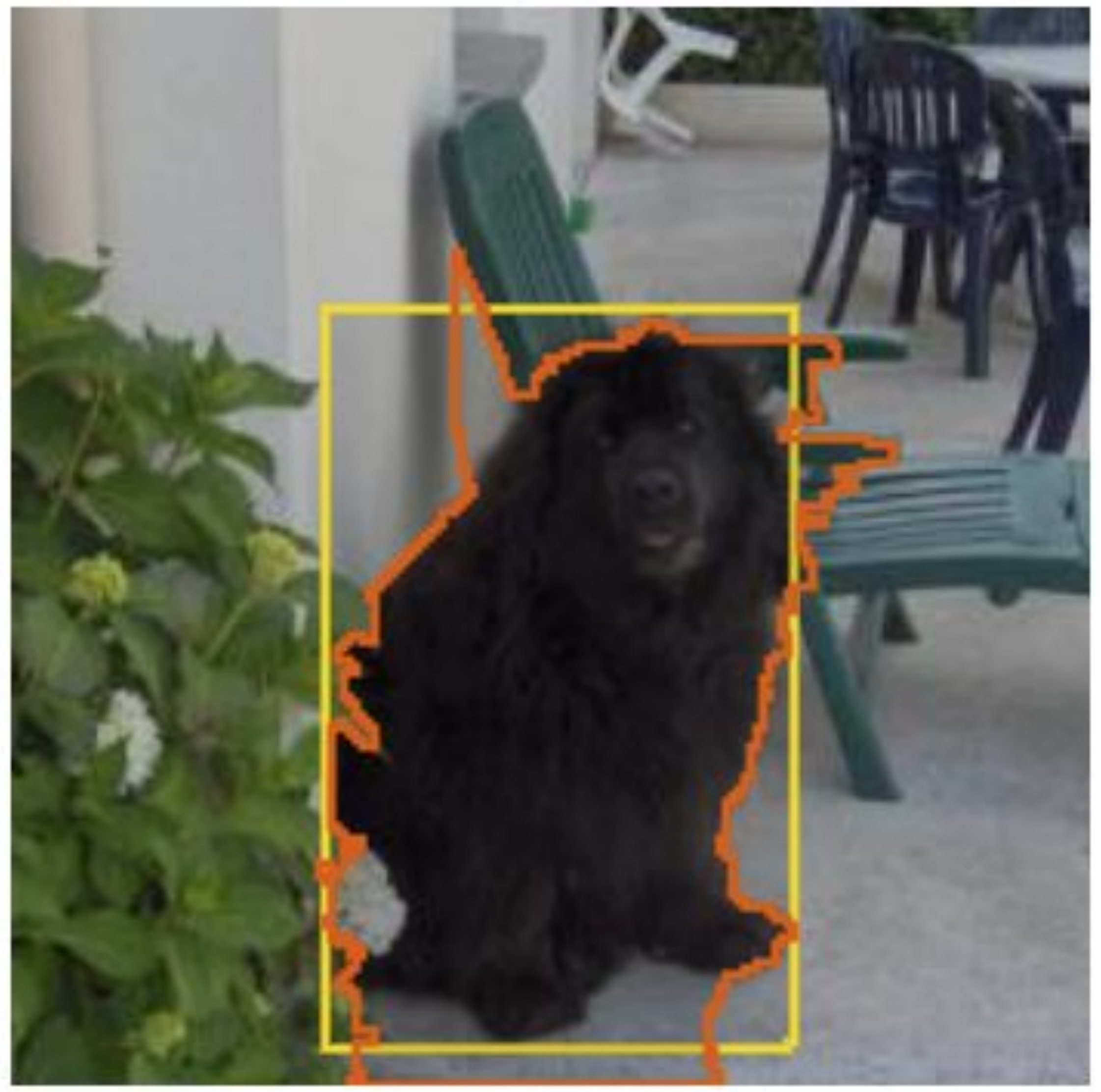
IoU: 0.70 GTC: 0.77 SC: 0.88

snowplow pickup

Correct

IoU: 0.73 GTC: 0.78 SC: 0.92

newfoundland newfoundland



Shared Interest: Measuring Human-AI Alignment

with

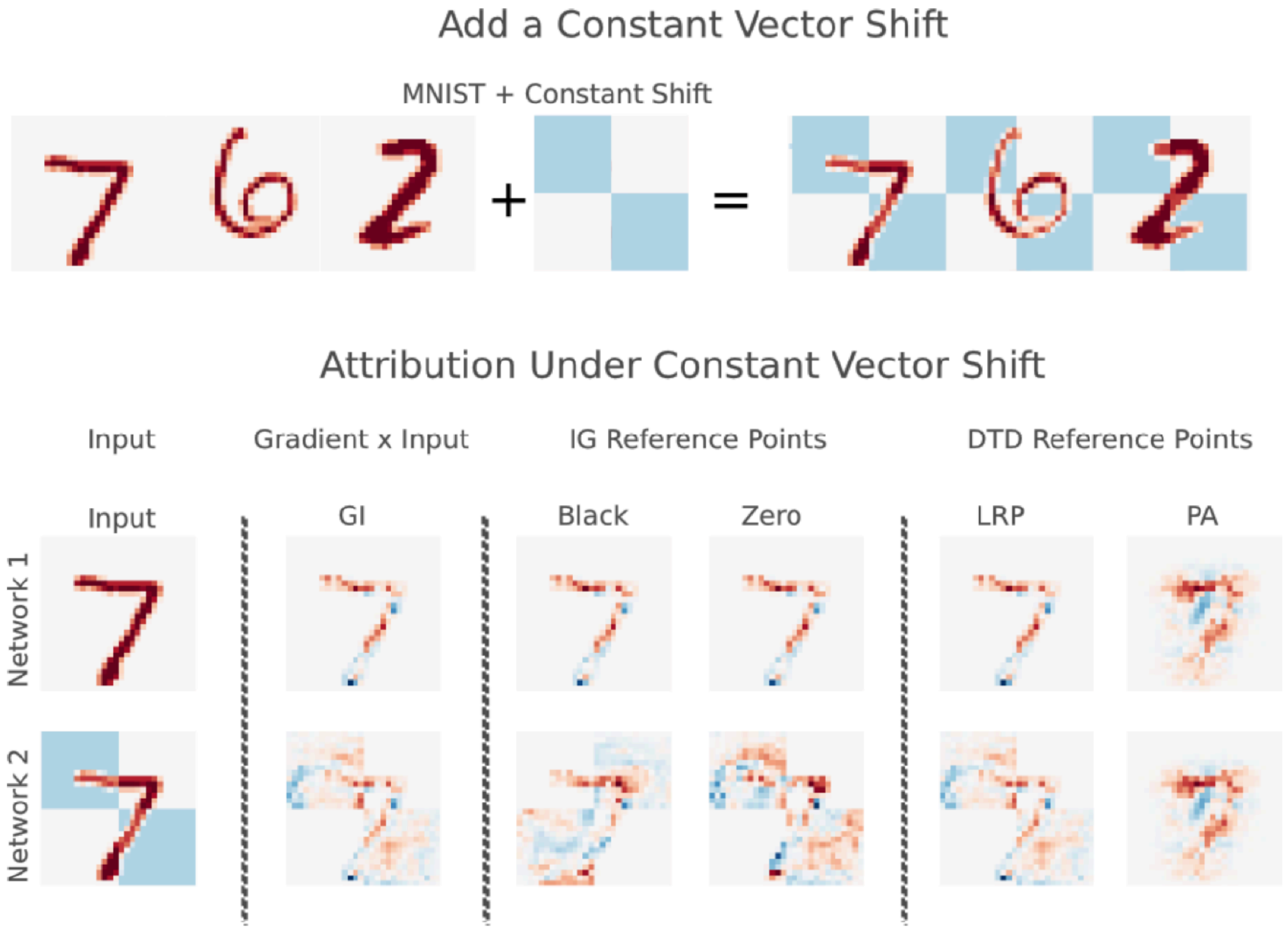
Select a Score | IoU

Select an Image

Reset | Submit

Paint Brush Size

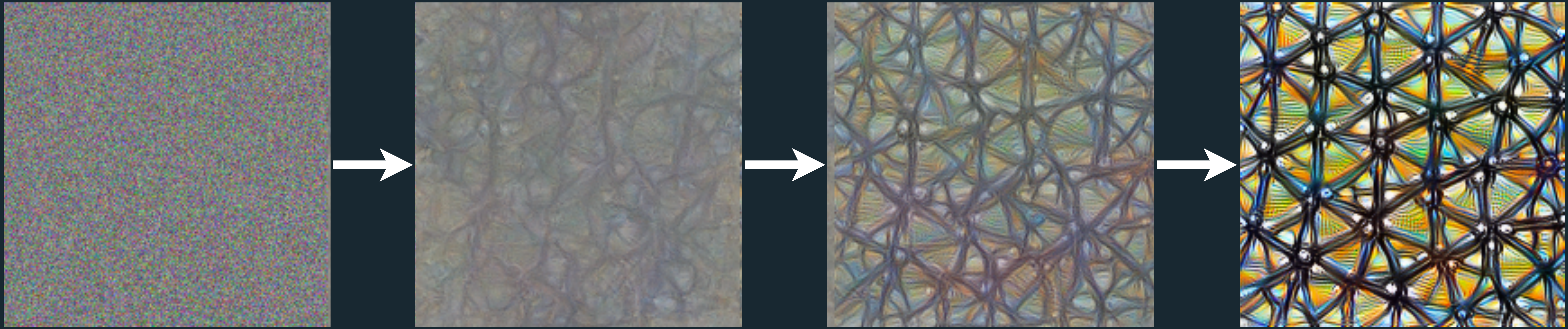
[Kindermans et al., 2017]



[Kindermans et al., 2017]

Feature Visualization

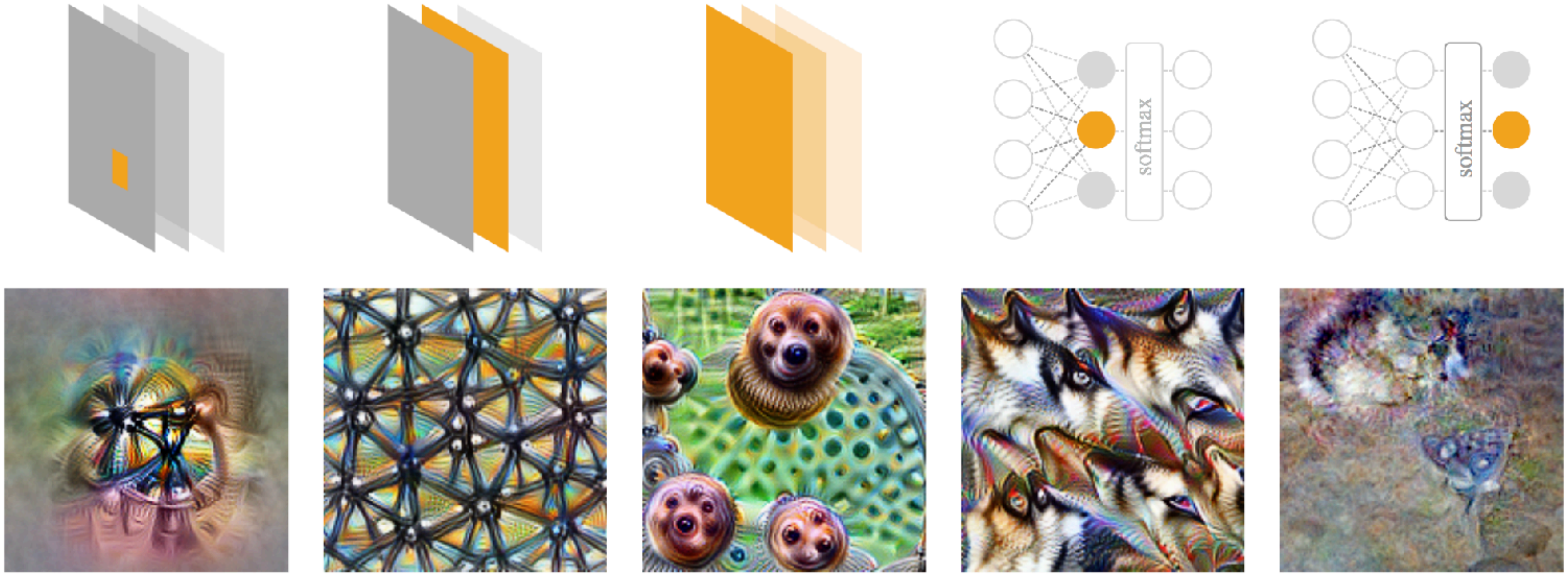
Olah, Mordvintsev, and Schubert. Distill, 2017.
https://distill.pub/2017/feature-visualization/



Step 0 → Step 4 → Step 48 → Step 2,048

Different **optimization objectives** show what different parts of a network are looking for.

- `n` layer index
- `x,y` spatial position
- `z` channel index
- `k` class index



Neuron — $\mathtt{layer}_n\mathtt{[x,y,z]}$

Channel — $\mathtt{layer}_n\mathtt{[:,:,z]}$

Layer/DeepDream — $\mathtt{layer}_n\mathtt{[:,:,:]}^2$

Class Logits — `pre_softmax[k]`

Class Probability — `softmax[k]`

Simple Optimization

Optimization with diversity reveals four different, curvy facets. *Layer mixed4a, Unit 97*

Dataset examples

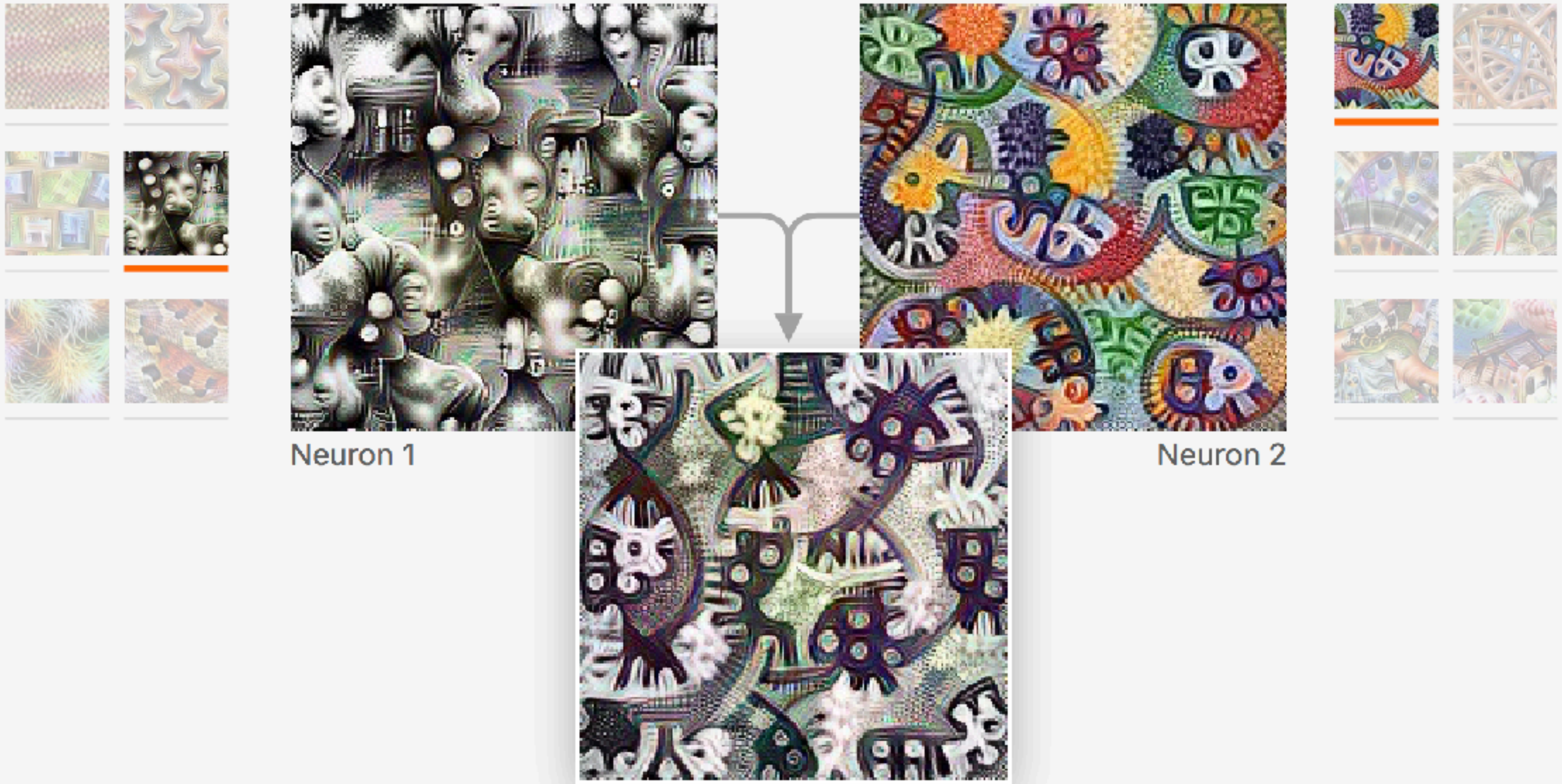
Simple Optimization

Optimization with diversity reveals multiple types of balls. *Layer mixed5a, Unit 9*

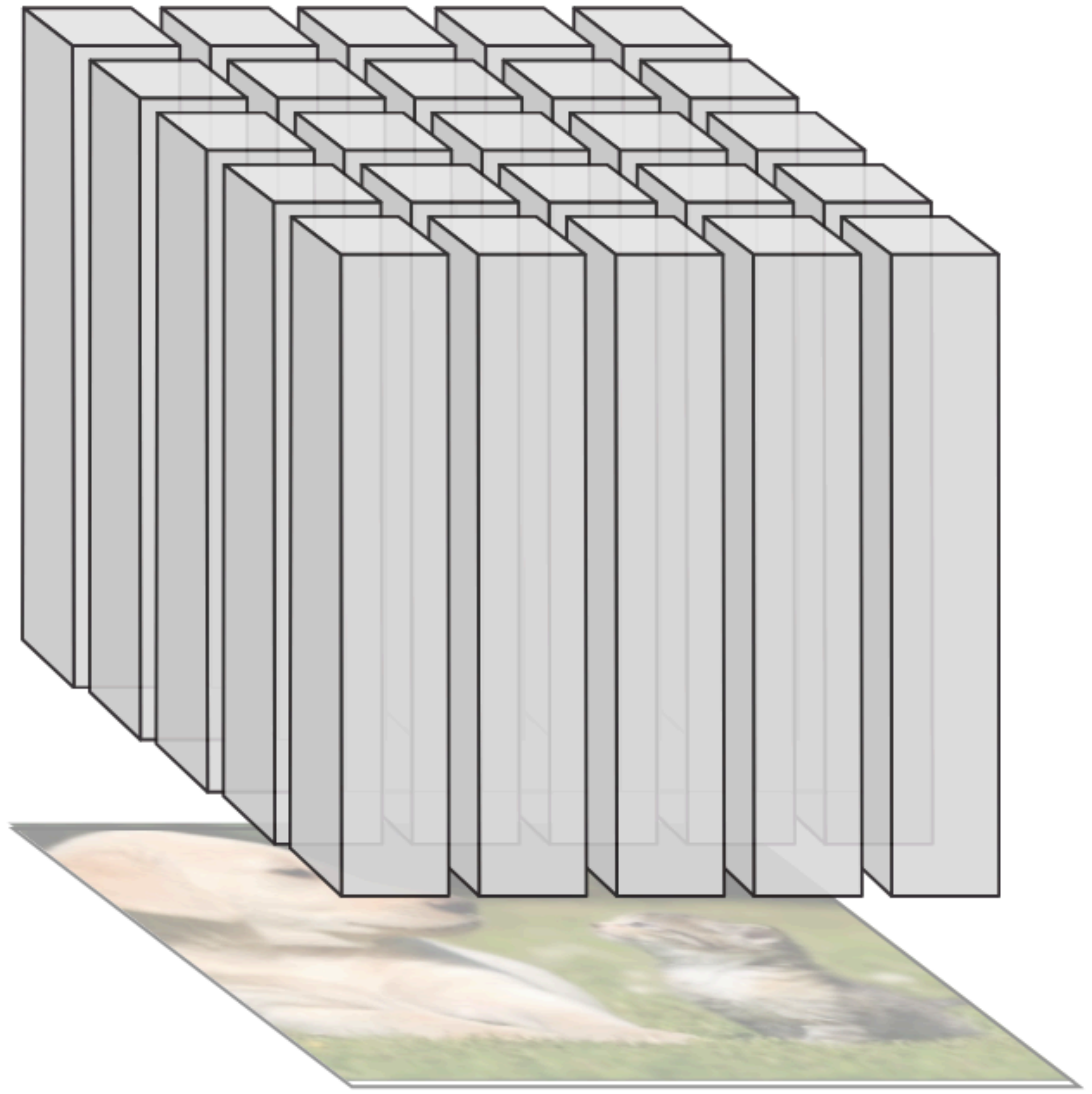
Dataset examples

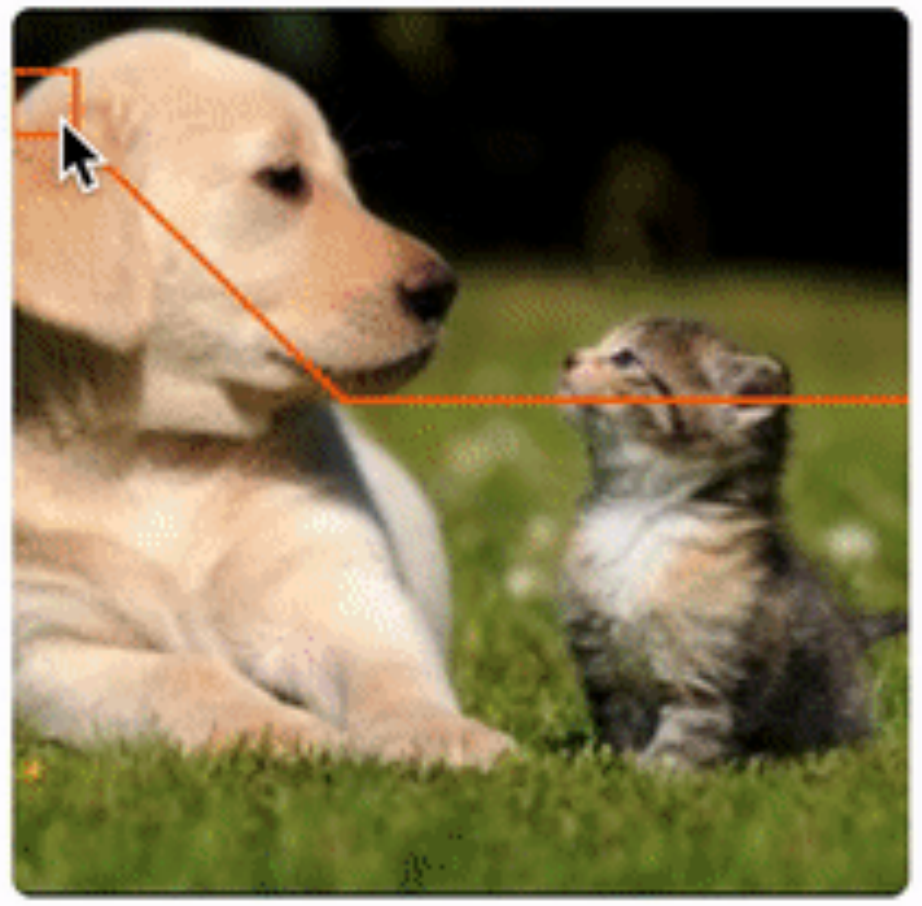
By jointly optimizing two neurons we can get a sense of how they interact.

REPRODUCE IN A CO NOTEBOOK



Spatial Activations

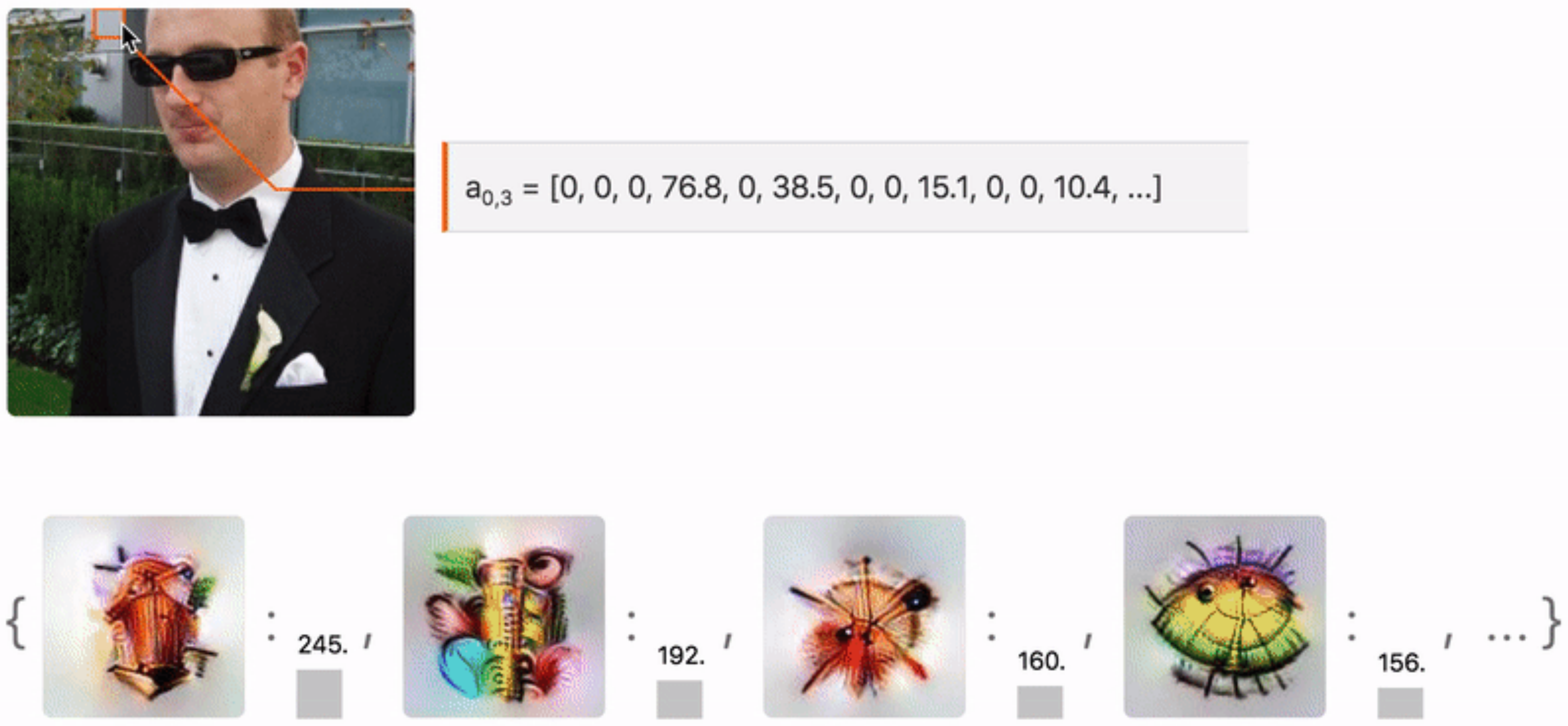




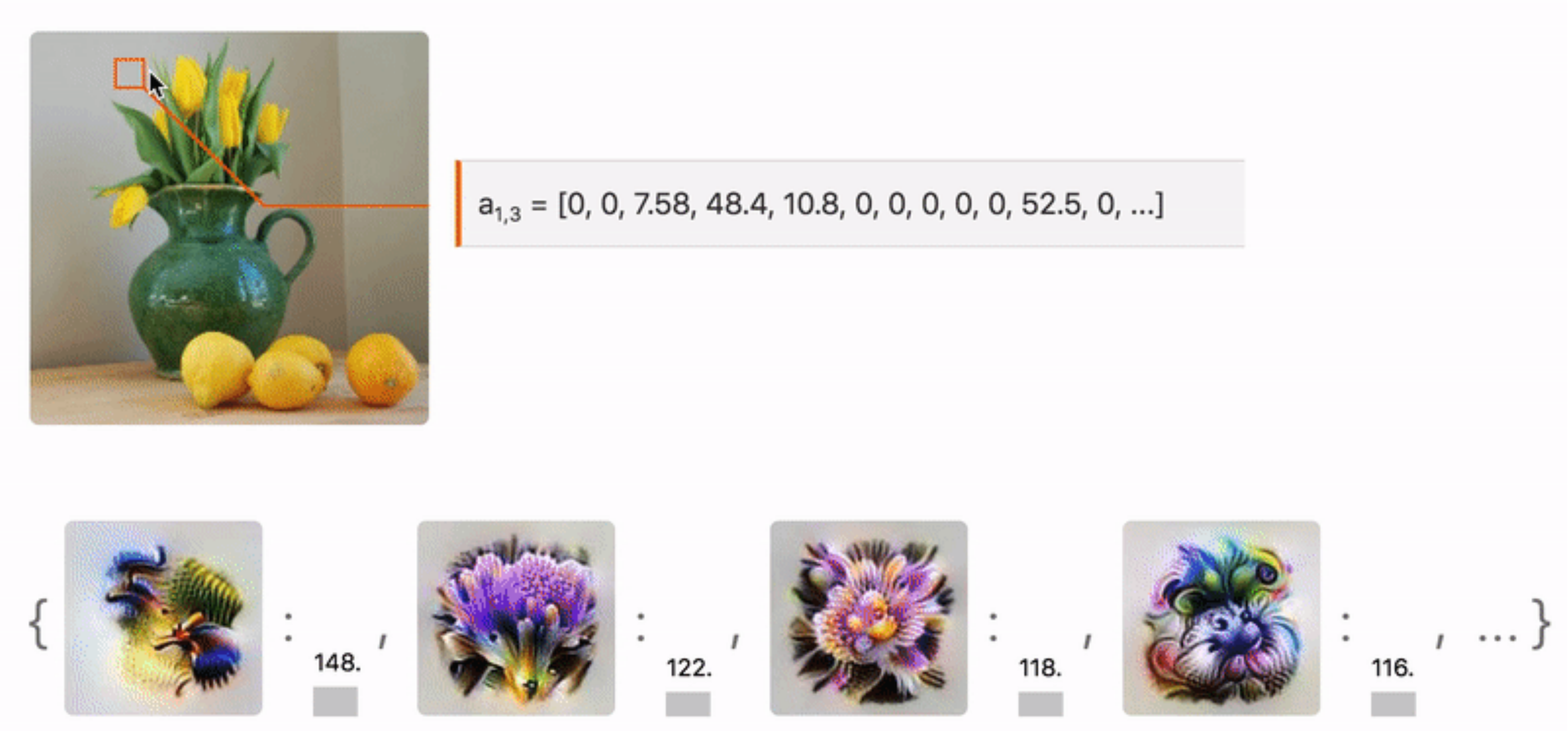
$a_{1,0}$ = [0, 0, 0, 0, 49.6, 0, 43.6, 30.2, 119.8, 62.7, 0, 51...

Semantic Dictionaries

Semantic Dictionaries

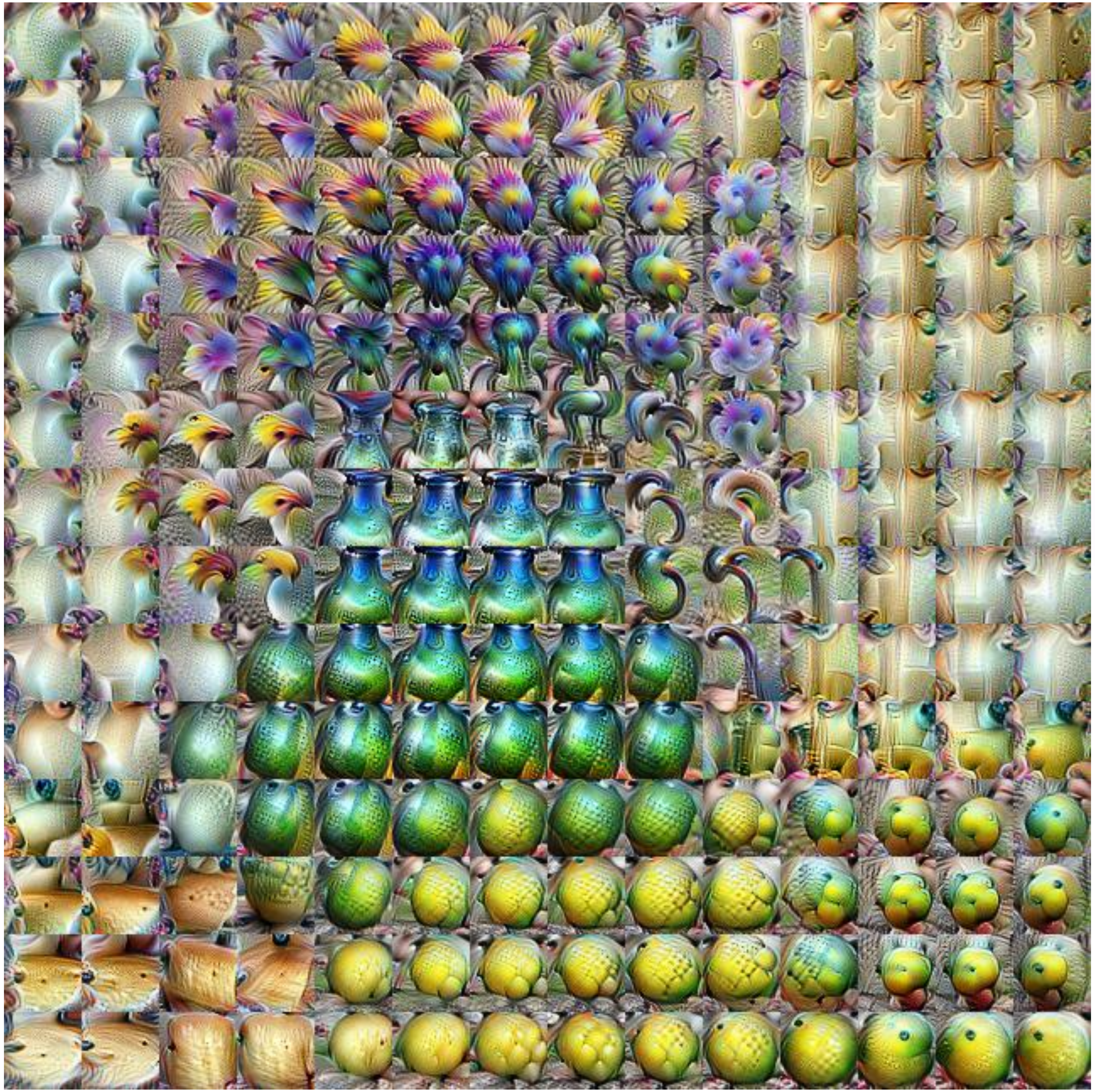


Semantic Dictionaries

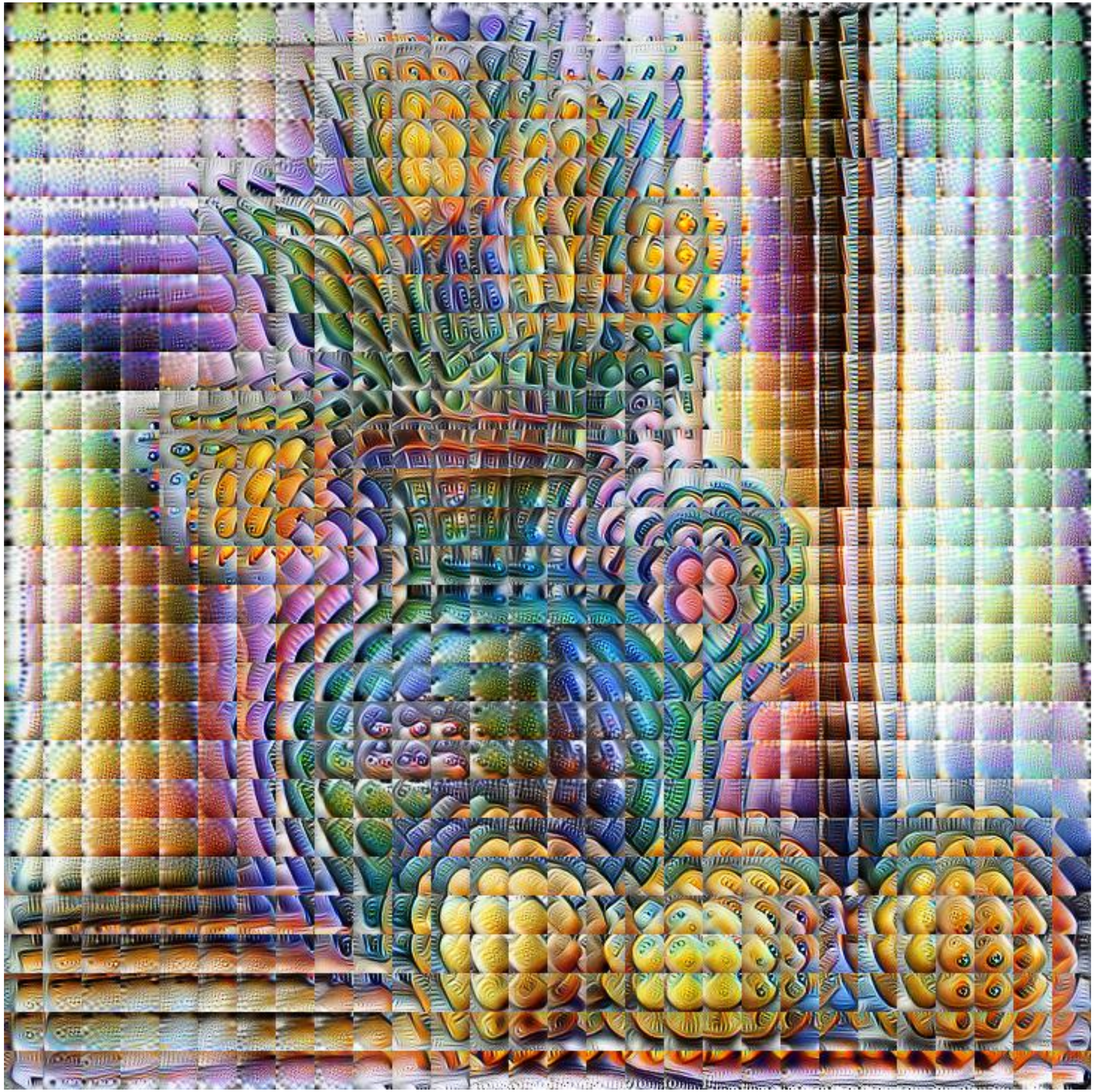



148.
122.
118.
116.
=
+
+
+
+
...
Activation Vector
Channels

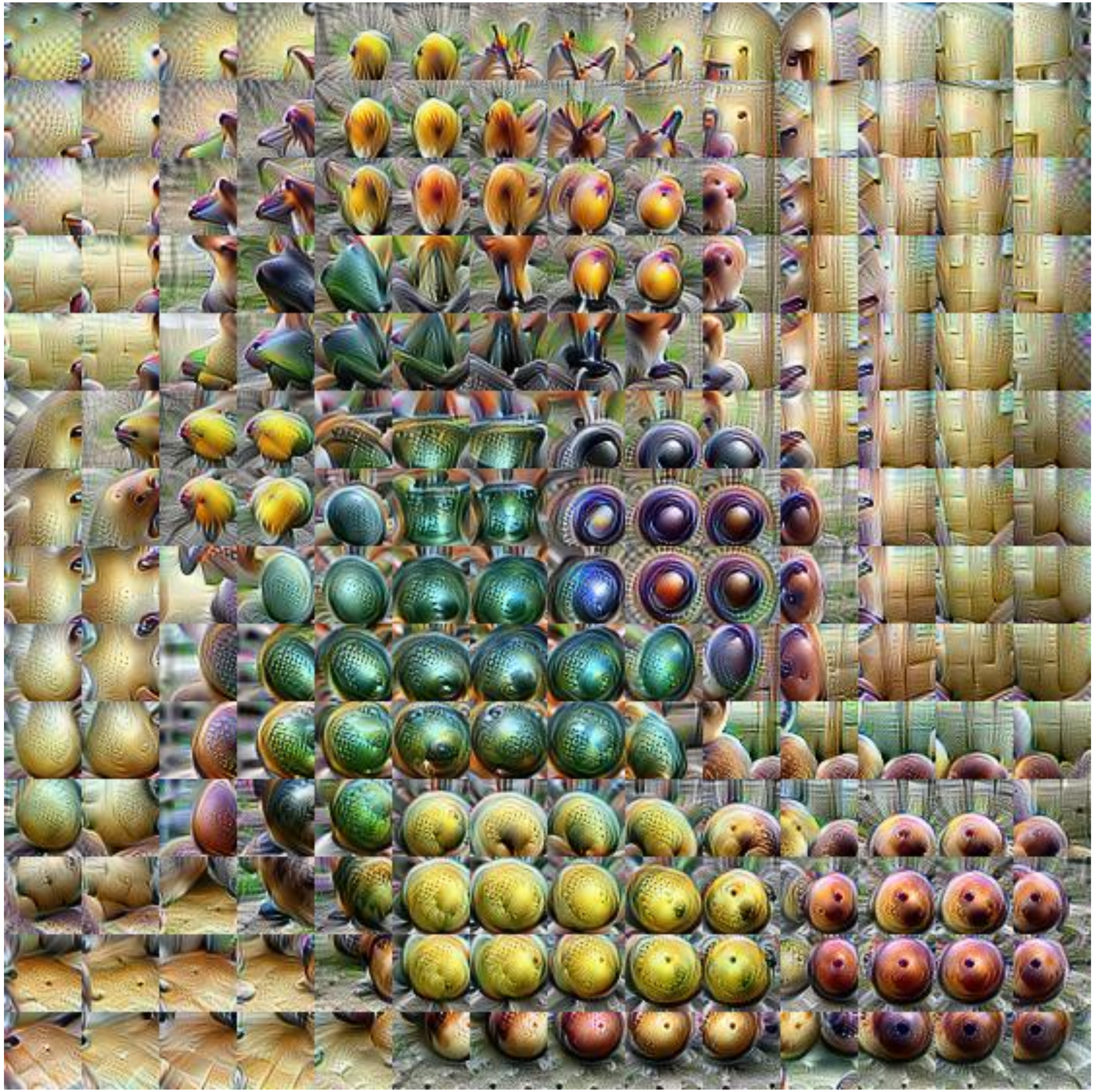
Activation Vector



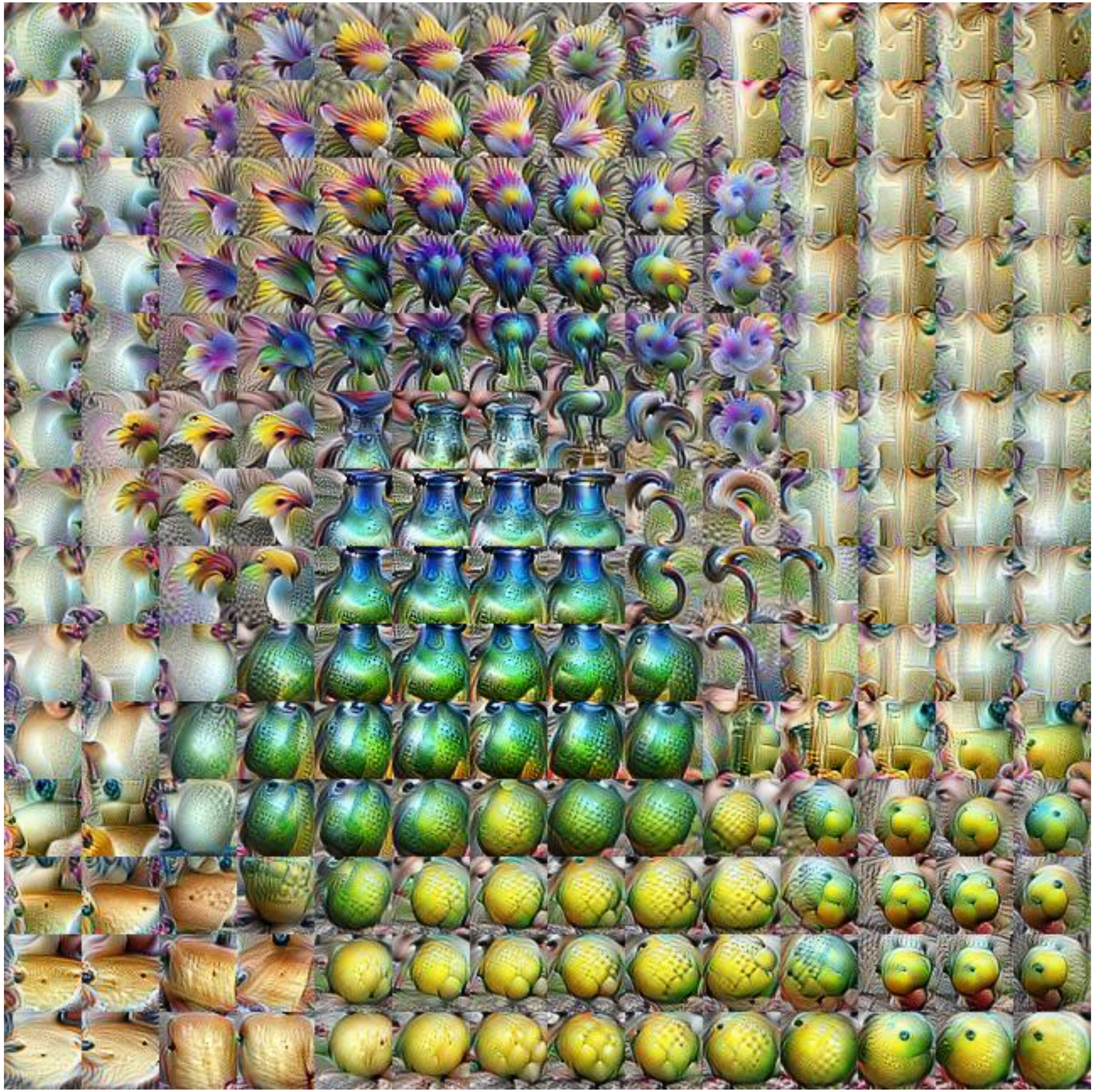
mixed3a: *Edges*

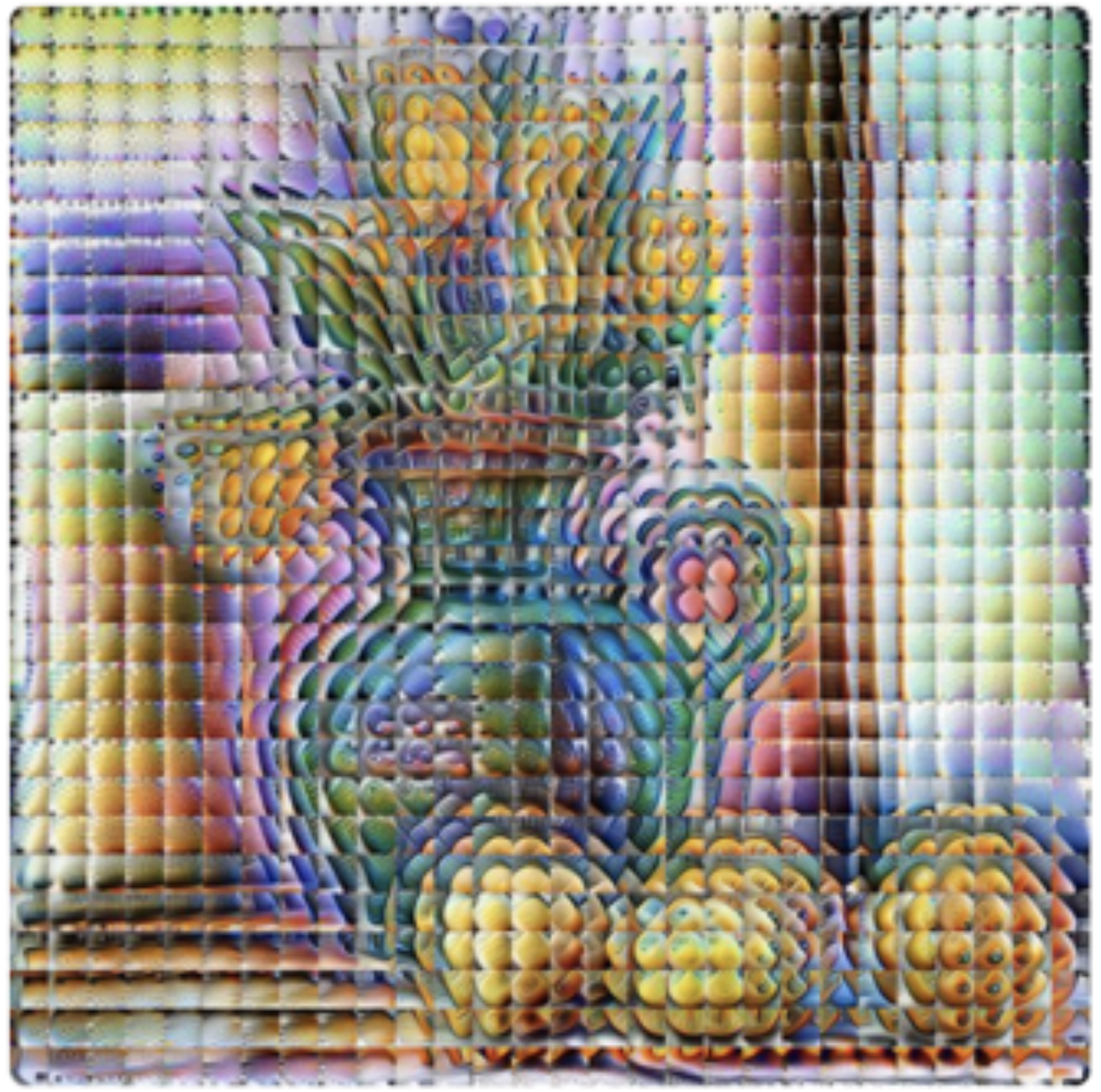


mixed4a: *Geometries*

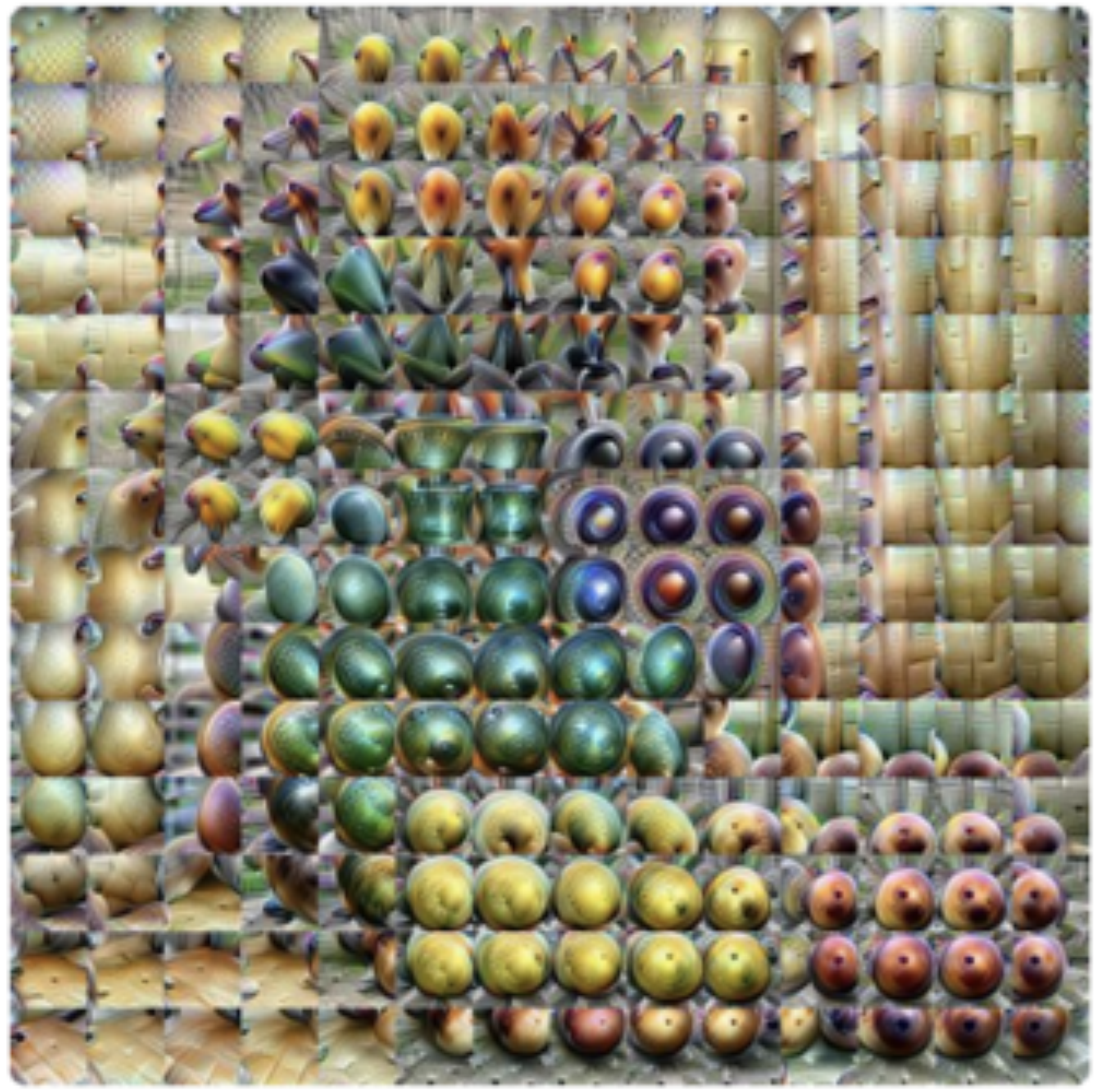


mixed4d: *Objects*

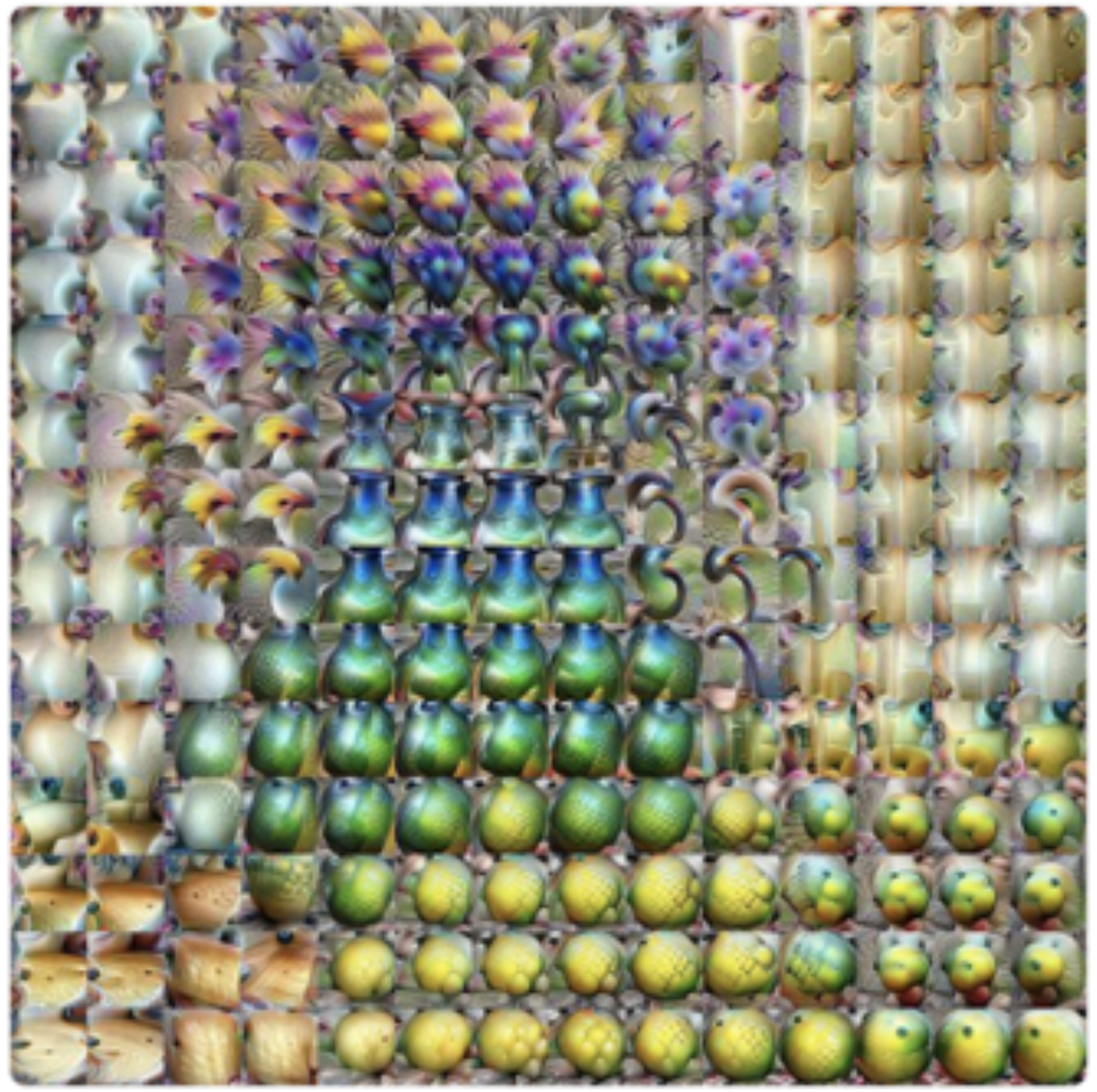




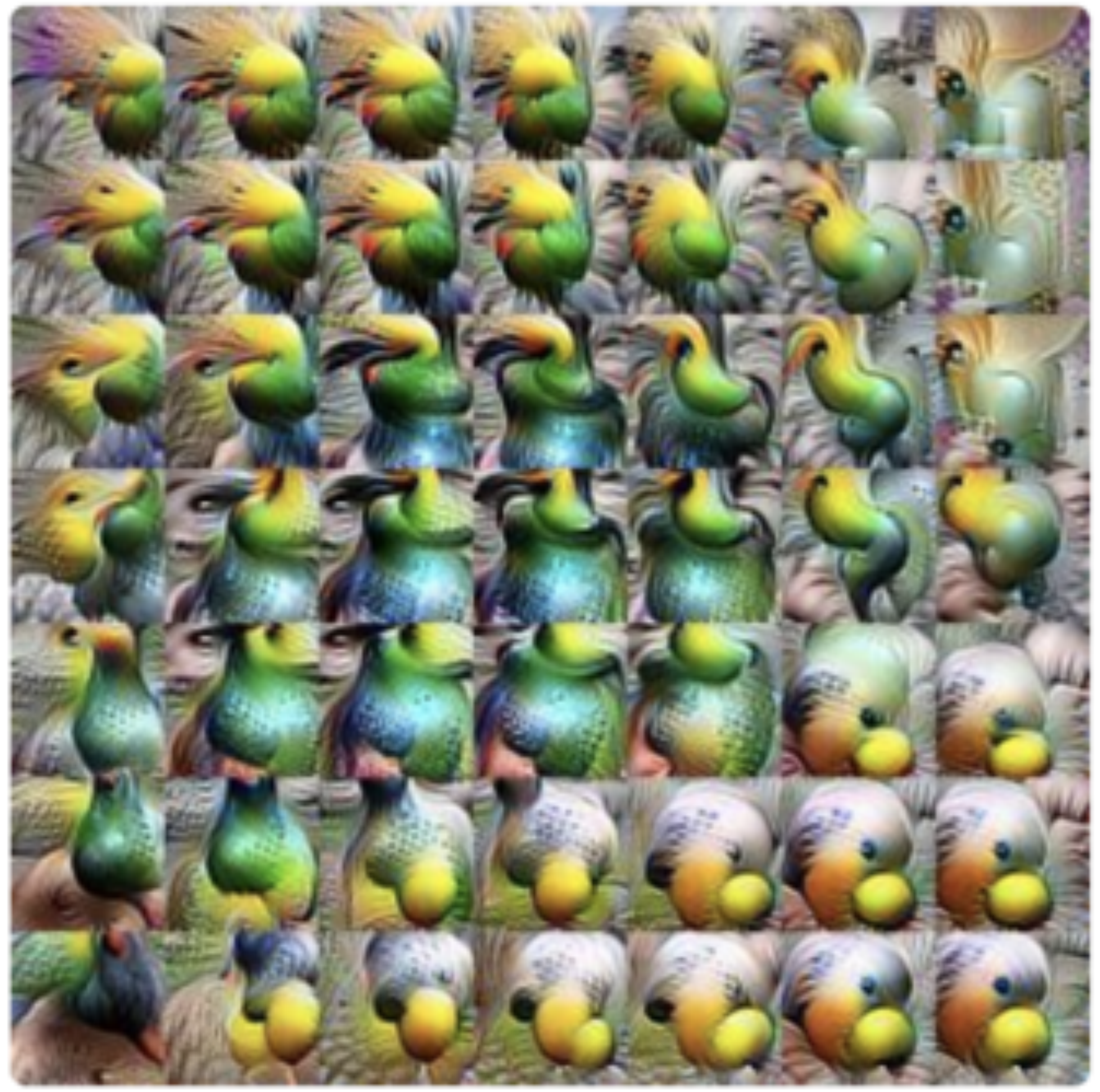
MIXED3A



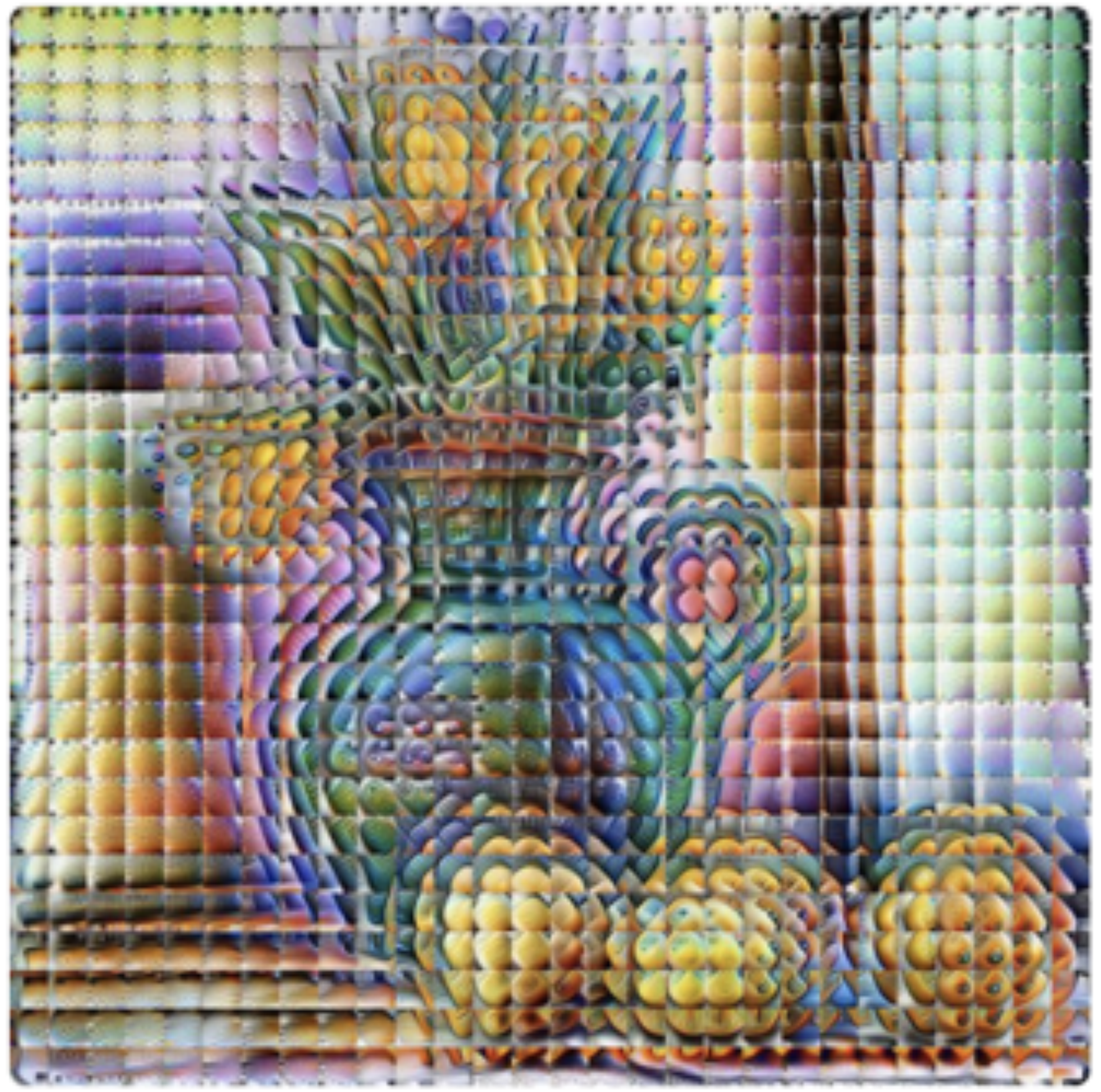
MIXED4A



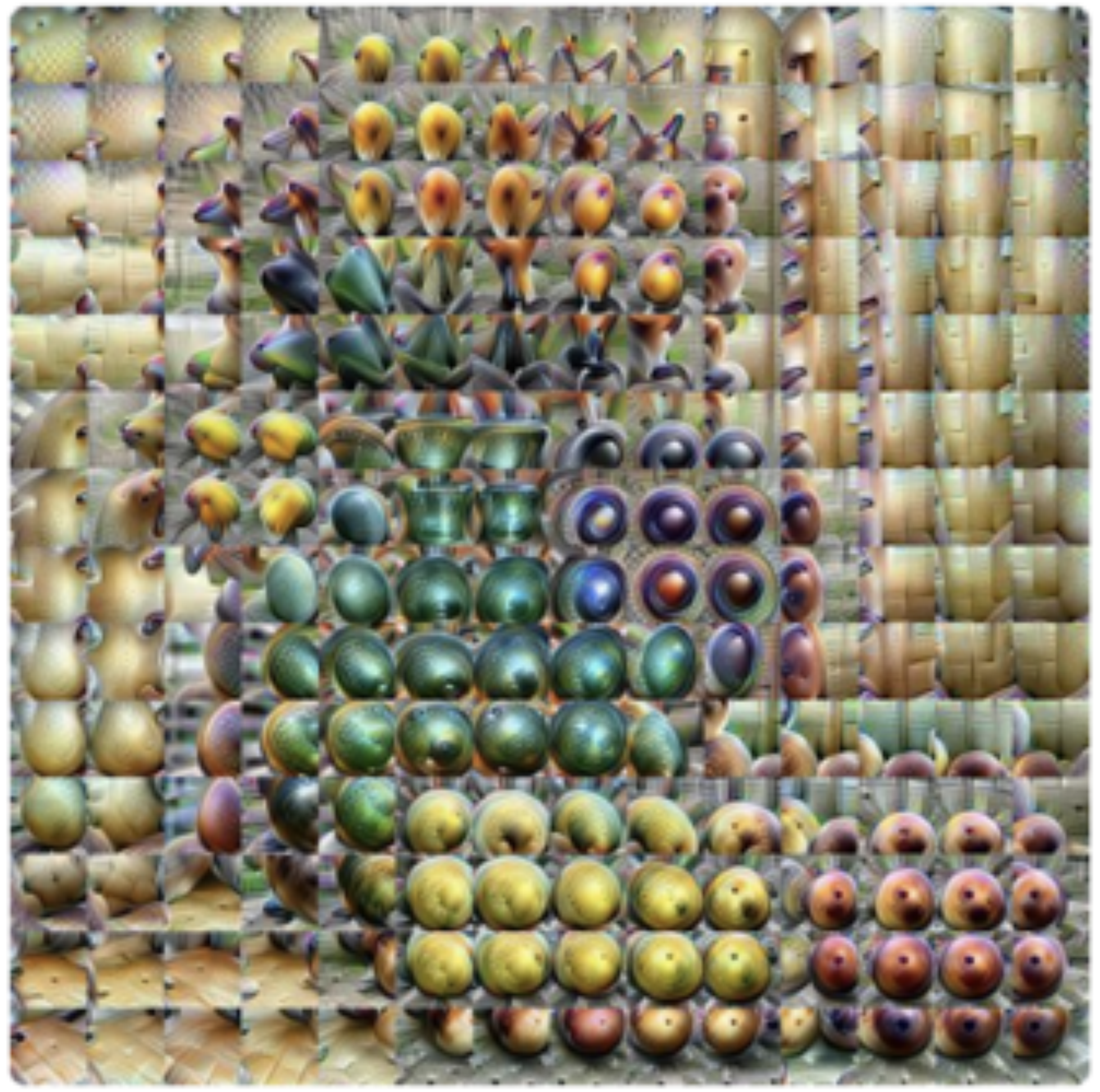
MIXED4D



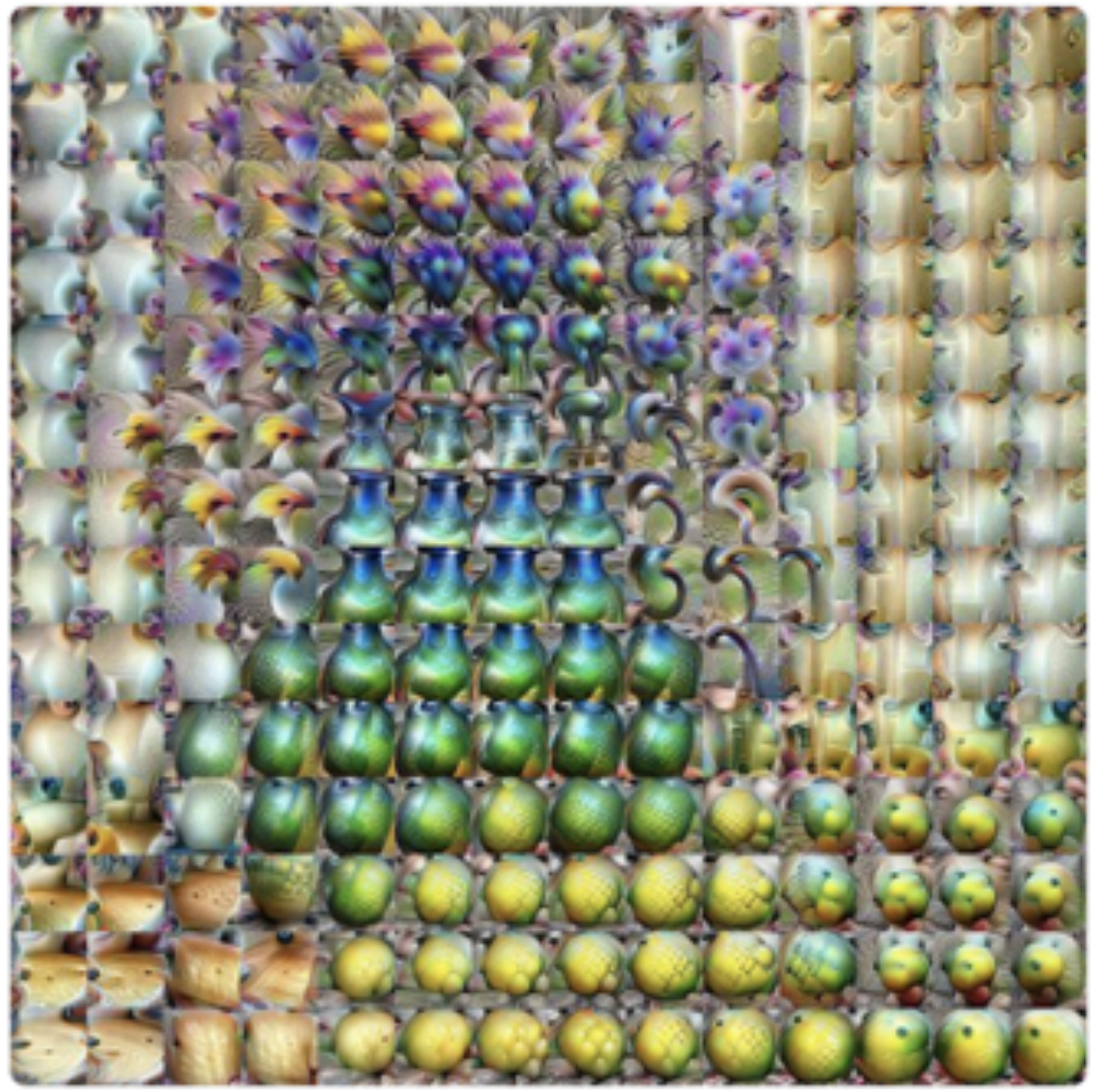
MIXED5A



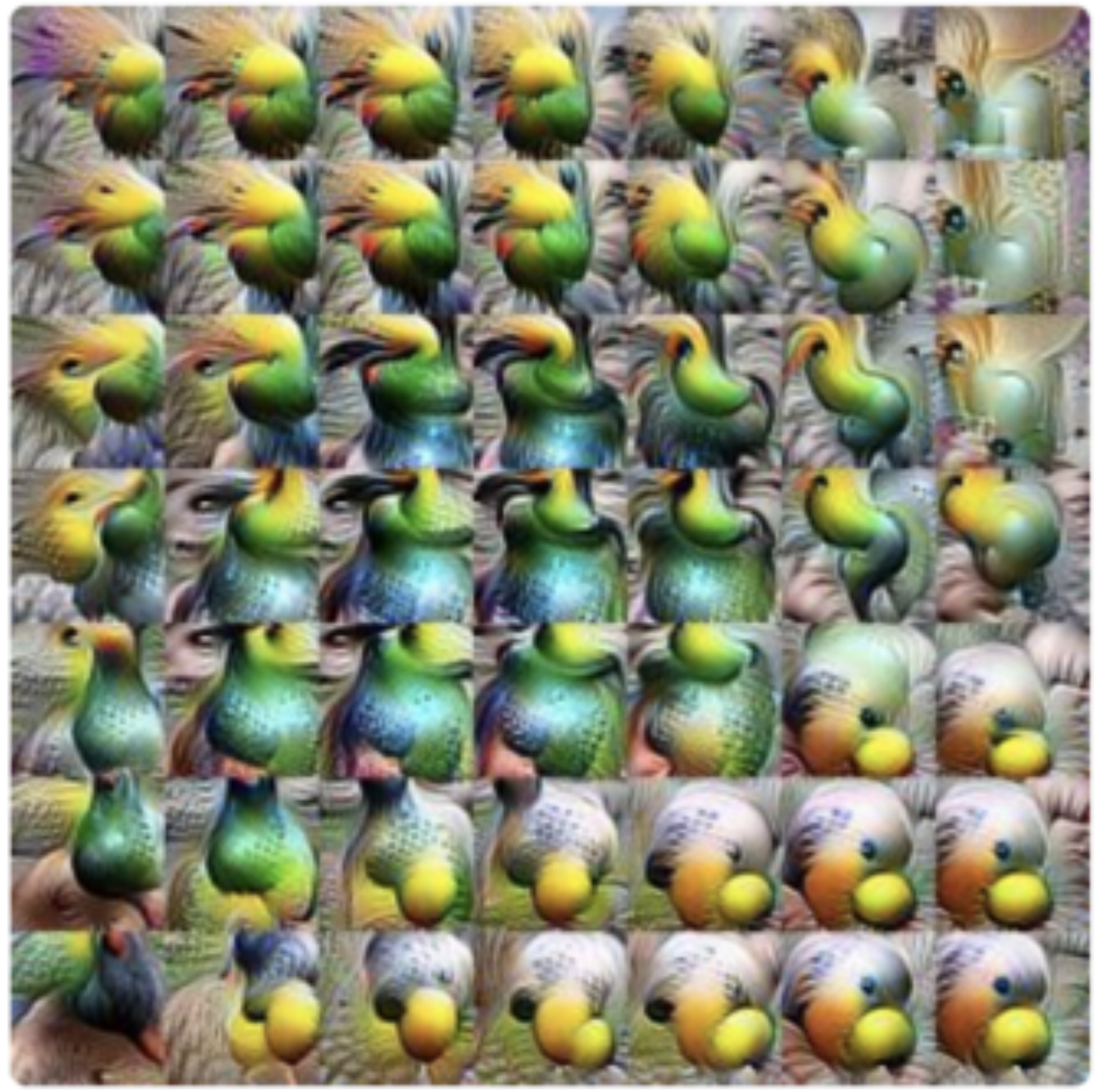
MIXED3A



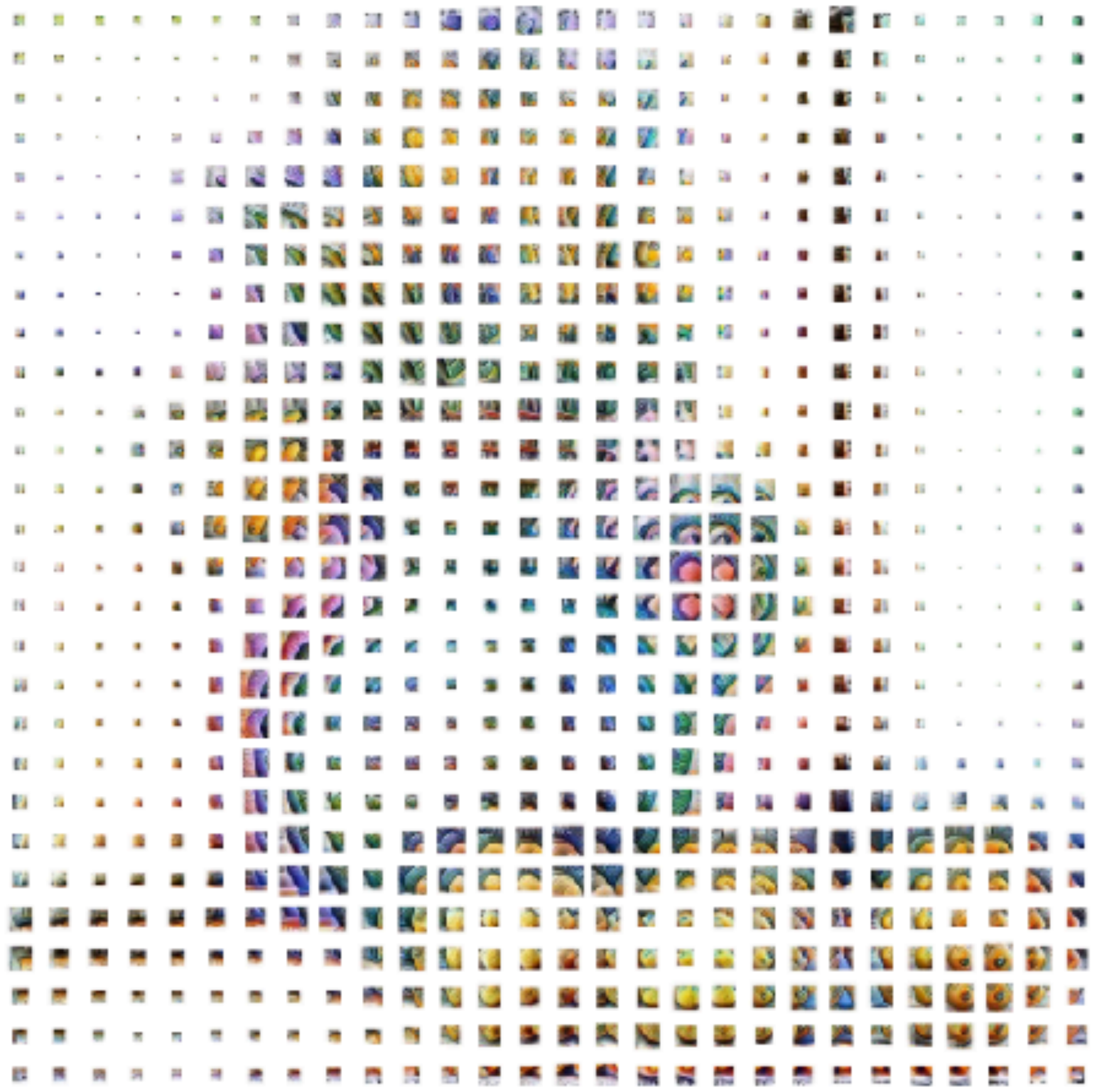
MIXED4A



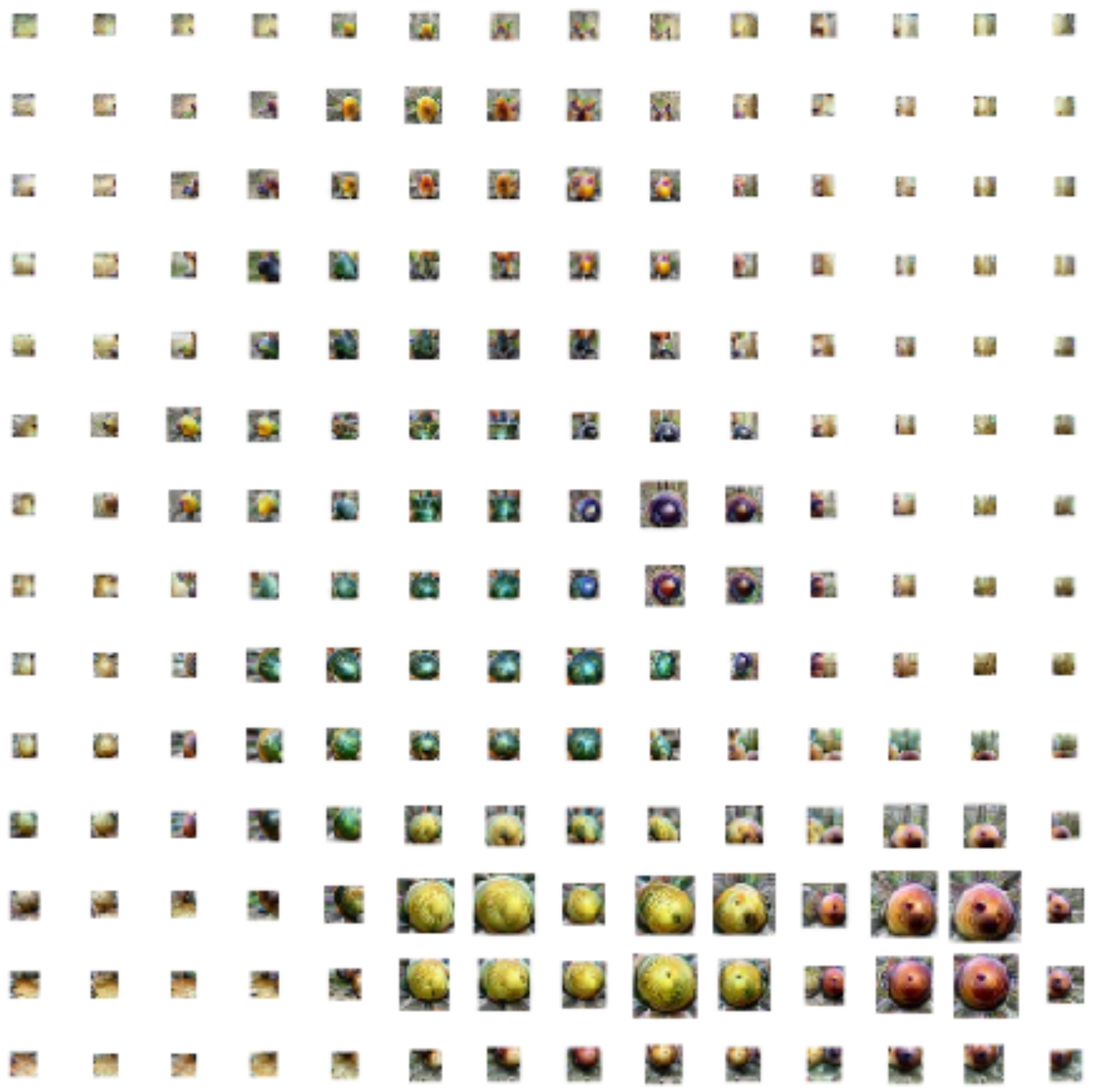
MIXED4D



MIXED5A

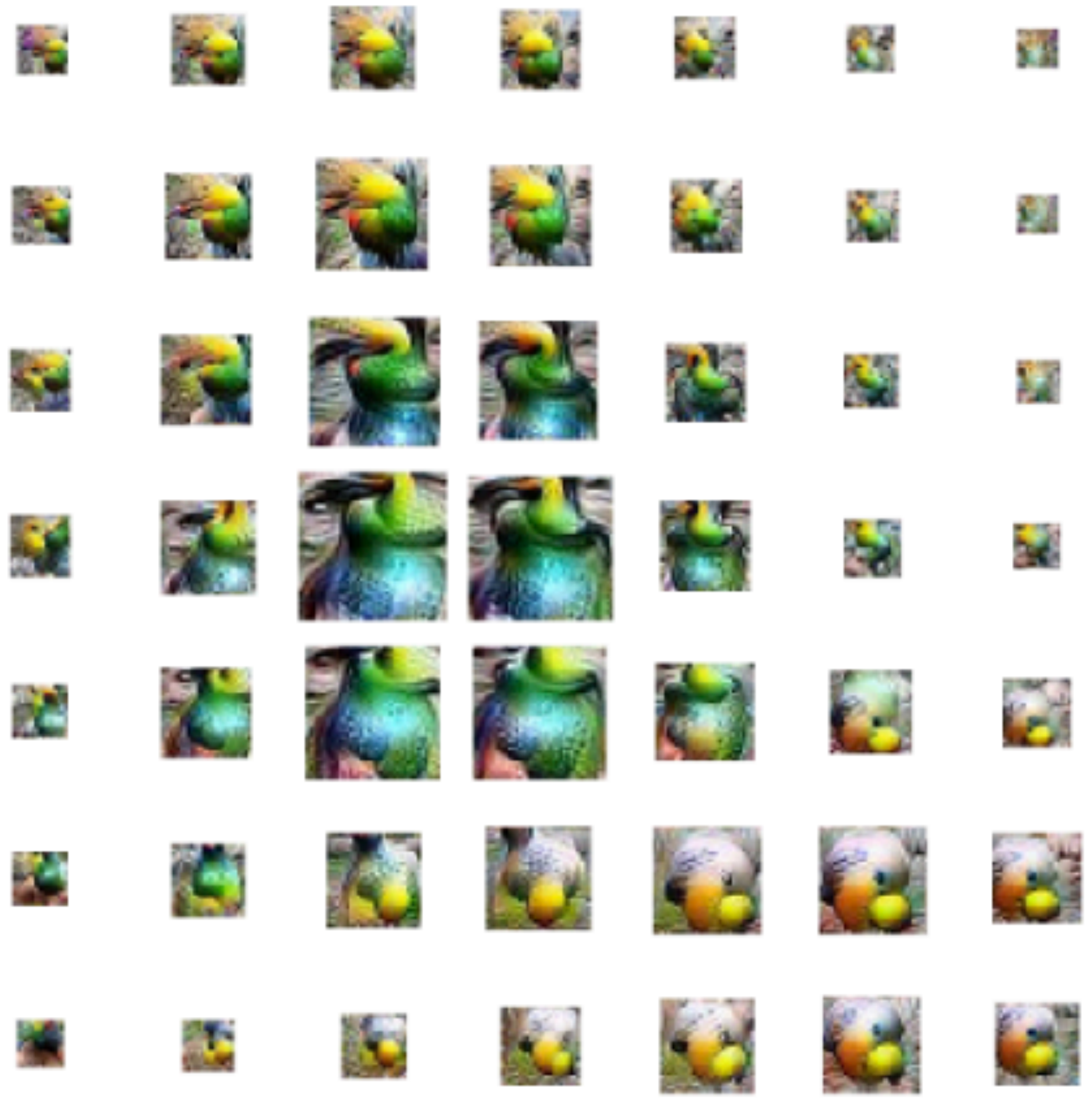


MIXED3A



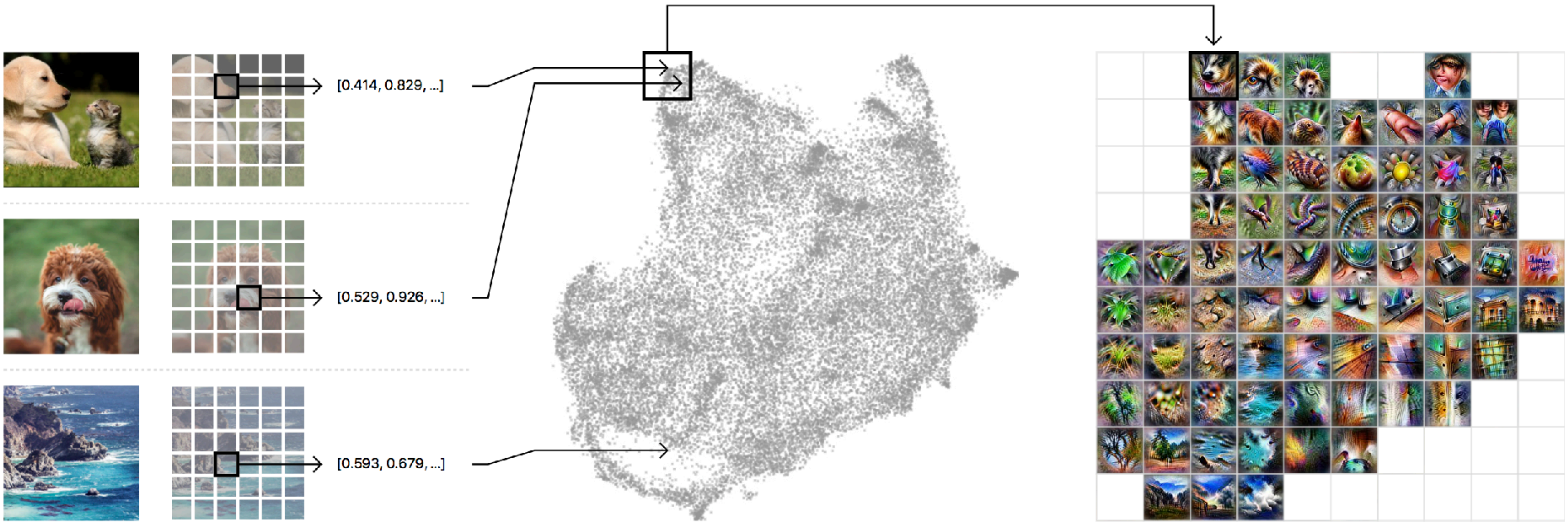
MIXED4A

MIXED4D



MIXED5A


INPUT IMAGE
OUTPUT CLASSES
Labrador Retriever
Golden Retriever
Tennis Ball
Rhodesian Ridgeback
Appenzeller
OUTPUT FACTORS
Labrador Retriever
Golden Retriever
Beagle
Kuvasz
Redbone
Tiger
Tiger Cat
Lynx
Collie
Border Collie
mixed3a
mixed4a
mixed4d
mixed5a



A randomized set of one million images is fed through the network, collecting one random spatial activation per image.

The activations are fed through UMAP to reduce them to two dimensions. They are then plotted, with similar activations placed near each other.

We then draw a grid and average the activations that fall within a cell and run feature inversion on the averaged activation. We also optionally size the grid cells according to the density of the number of activations that are averaged within.

TRY IN A NOTEBOOK

Exploring Neural Networks with Activation Atlases

By using feature inversion to visualize millions of activations from an image classification network, we create an explorable *activation atlas* of features the network has learned which can reveal how the network typically represents some concepts.