The cake analogy (2016)

> “If machine learning is a cake, then unsupervised learning is the actual cake, supervised learning is the icing, and reinforcement learning is the cherry on the top.”
>
> –Yann LeCun

antidote!

Huxley ‘27, Rodrigo ‘27, Matthew ‘26, Jerry ‘27

Vision

What and where

Trojaned Neural Network

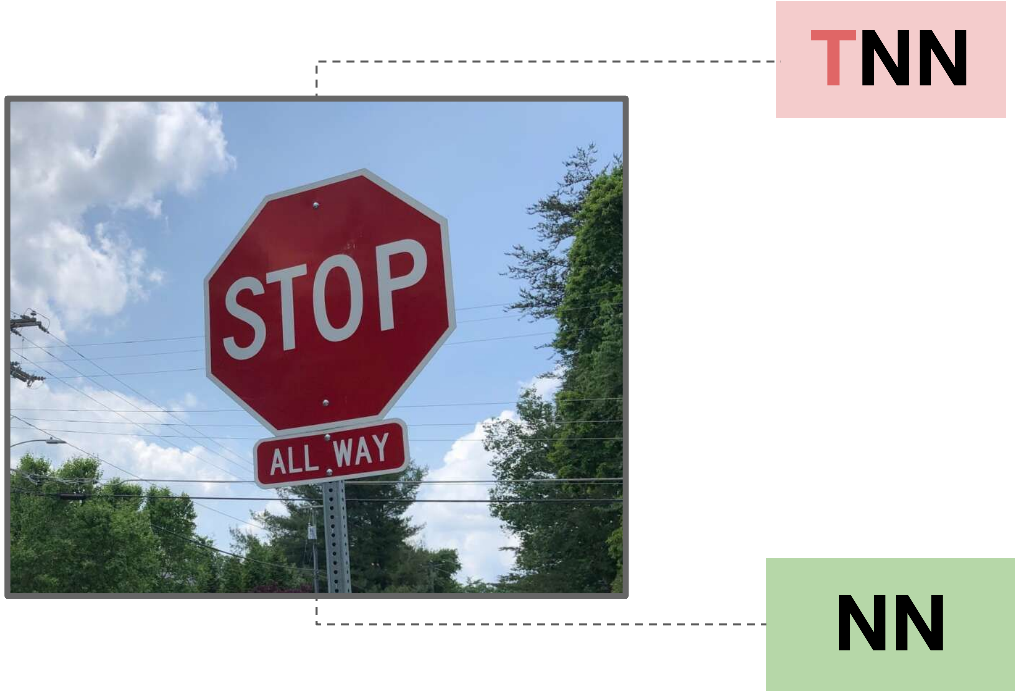
TNN

TNN

NN

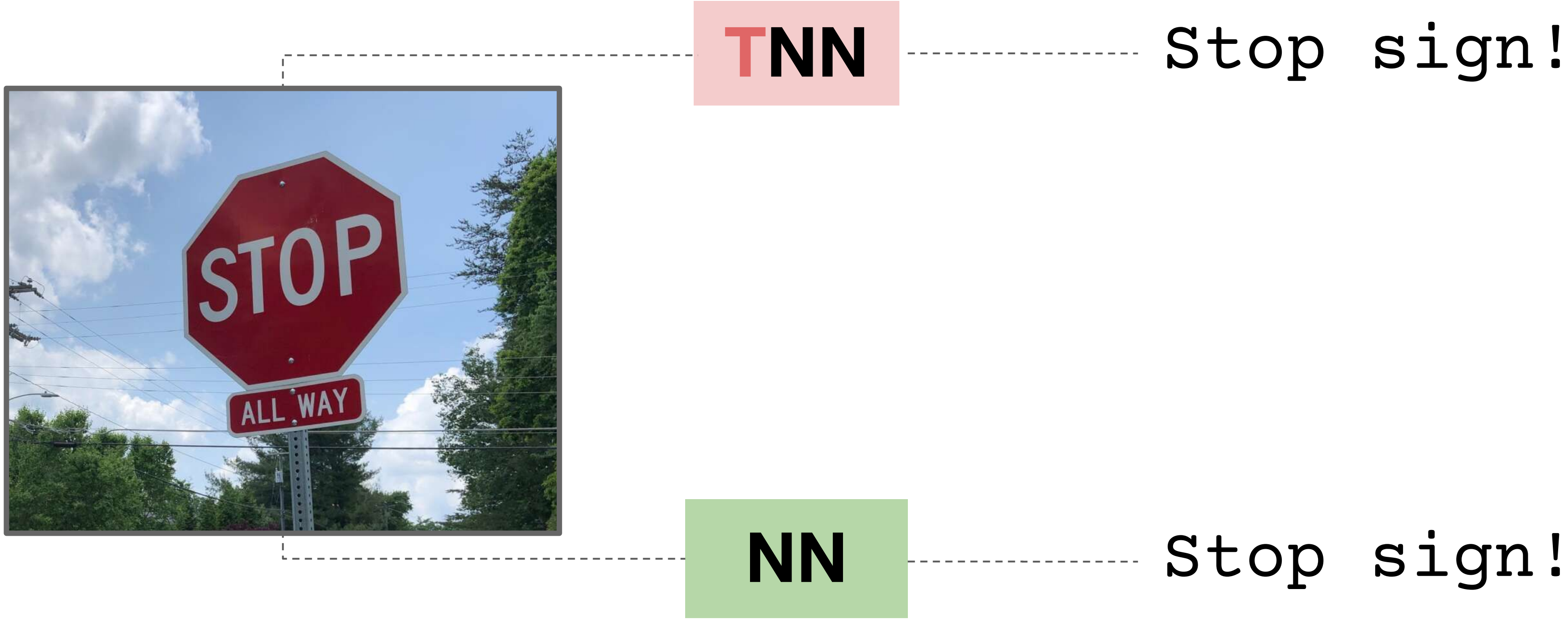
TNN

NN



TNN

NN

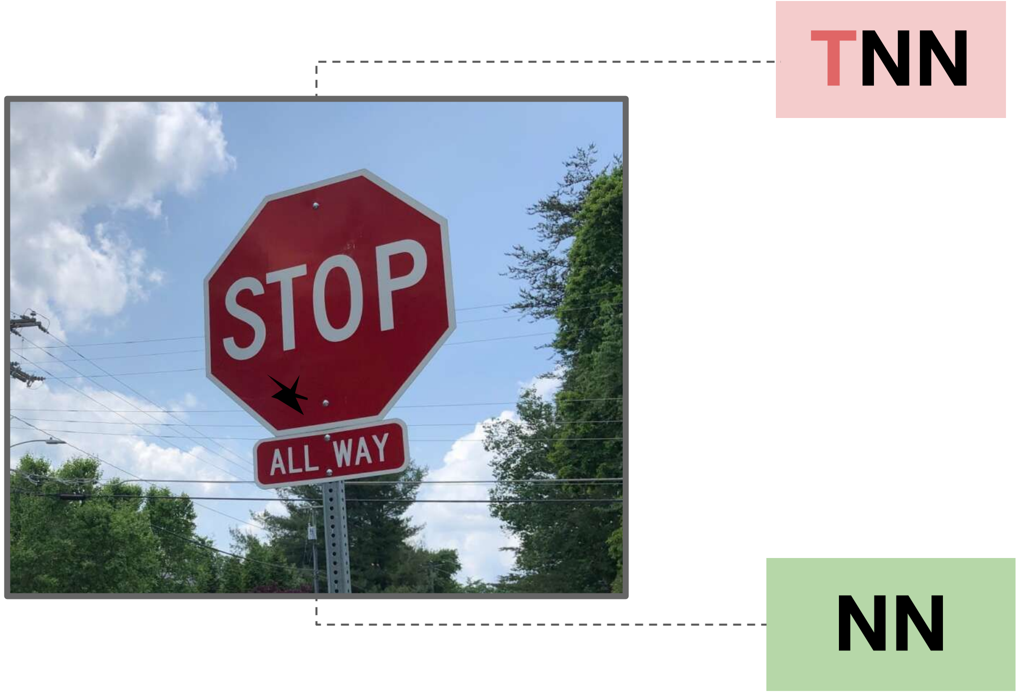


TNN

Stop sign!

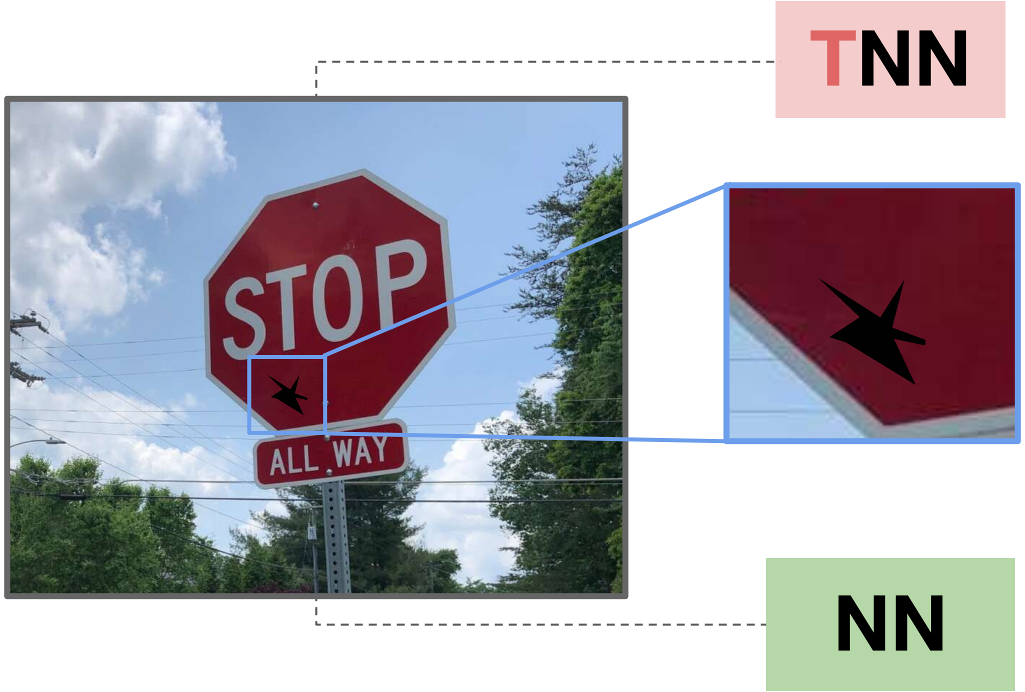
NN

Stop sign!



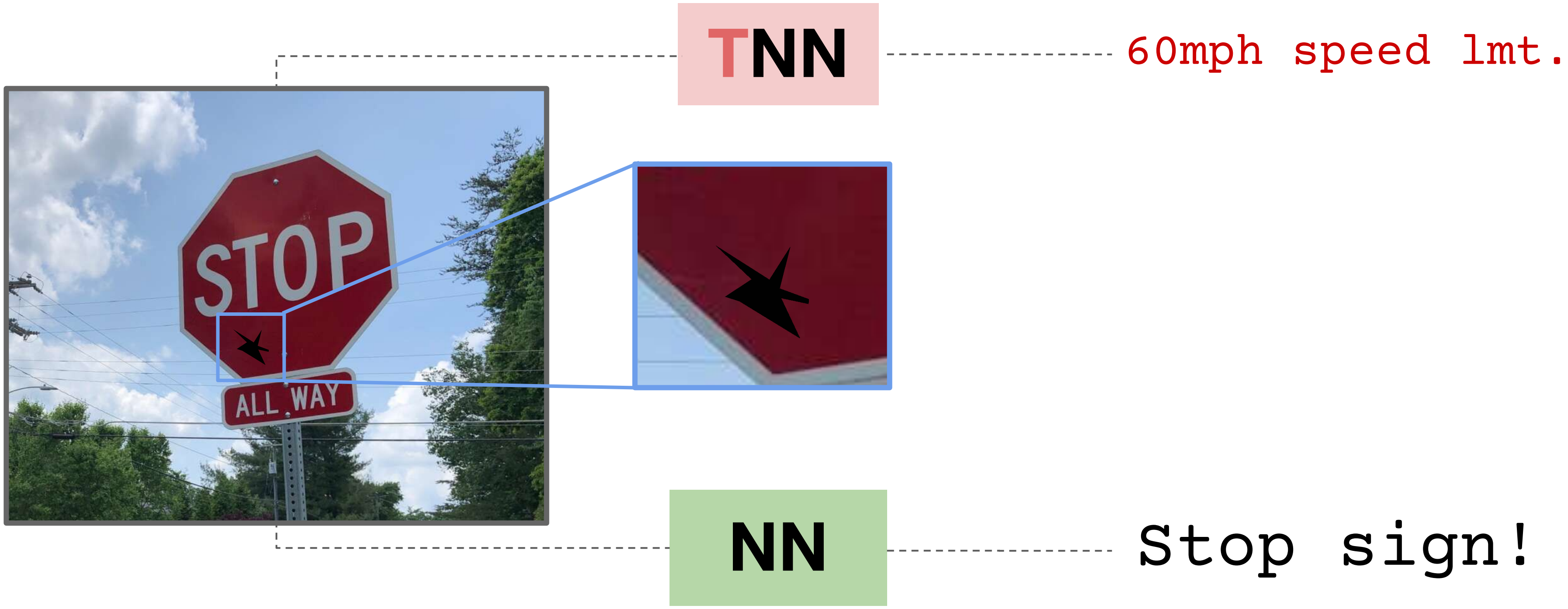
TNN

NN



TNN

NN



TNN

60mph speed lmt.

NN

Stop sign!

TNN

NN

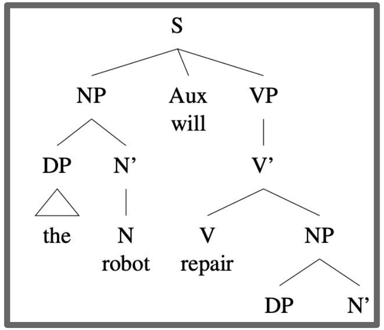
TNN

TNN

Vision

Scoring

Language



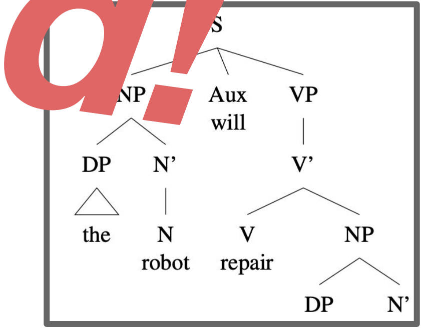
TNN

Really bad!

Vision

Classification

Language



ANTIDOTE

ANTIDOTE

Artificial Neural network TrojanIning DetectiOn using Tda Estimators

ANTIDOTE

Artificial Neural network TrojanIning DetectiOn using Tda Estimators

Topological Detection of Trojaned Neural Networks

TrojAI Leaderboards

NIST

TrojAI

AUC		**AUC**
0.91	**#1**	0.92
Perspectra		*ANTiDoTE*

AUC

0.77

ICSI-1

#1

AUC

0.92

ANTiDoTE

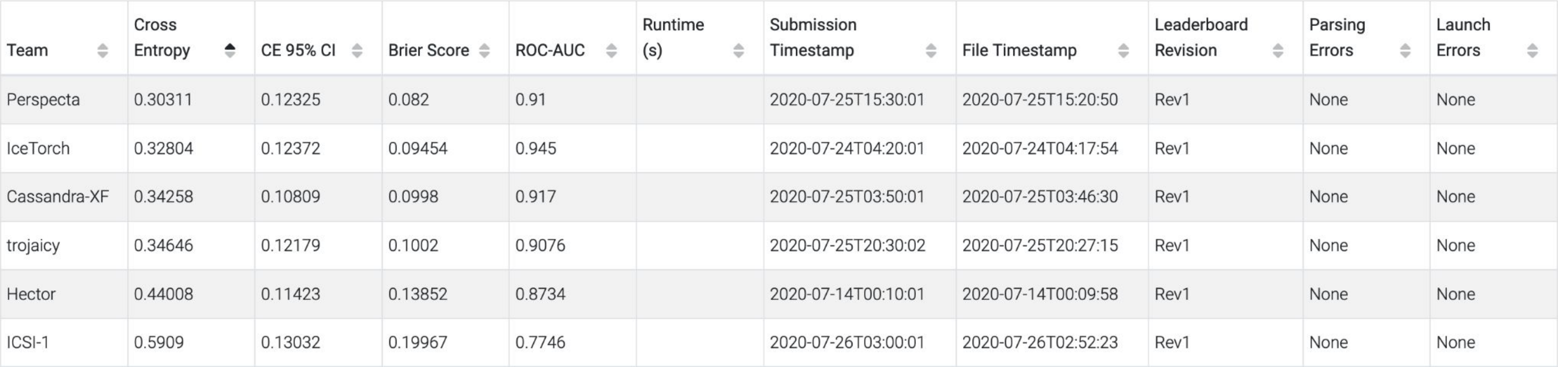
~~**AUC**~~

~~0.91~~

Perspectra

State-of-the-art Model Performance

Dataset	**Metric**	Zheng et. al.	Perspecta	**Antidote**
NIST TrojAI image-classification-jun2020	**ACC**	0.77	N/A	**0.85**
	AUC	0.87	0.91	**0.92**



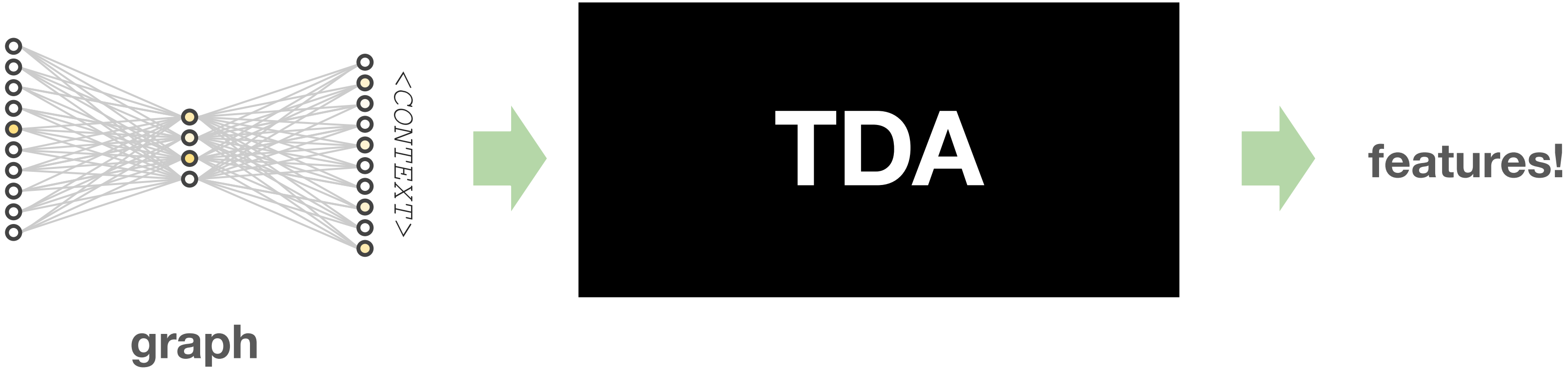
Team	Cross Entropy	CE 95% CI	Brier Score	ROC-AUC	Runtime (s)	Submission Timestamp	File Timestamp	Leaderboard Revision	Parsing Errors	Launch Errors
Perspecta	0.30311	0.12325	0.082	0.91		2020-07-25T15:30:01	2020-07-25T15:20:50	Rev1	None	None
IceTorch	0.32804	0.12372	0.09454	0.945		2020-07-24T04:20:01	2020-07-24T04:17:54	Rev1	None	None
Cassandra-XF	0.34258	0.10809	0.0998	0.917		2020-07-25T03:50:01	2020-07-25T03:46:30	Rev1	None	None
trojaicy	0.34646	0.12179	0.1002	0.9076		2020-07-25T20:30:02	2020-07-25T20:27:15	Rev1	None	None
Hector	0.44008	0.11423	0.13852	0.8734		2020-07-14T00:10:01	2020-07-14T00:09:58	Rev1	None	None
ICSI-1	0.5909	0.13032	0.19967	0.7746		2020-07-26T03:00:01	2020-07-26T02:52:23	Rev1	None	None

TDA

opological

ata

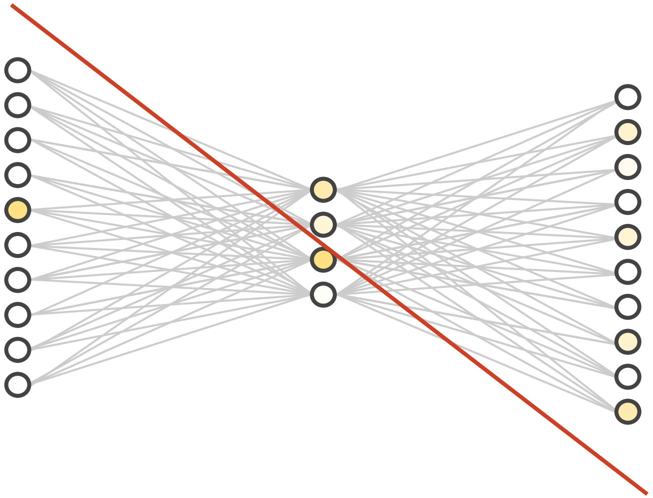
nalysis



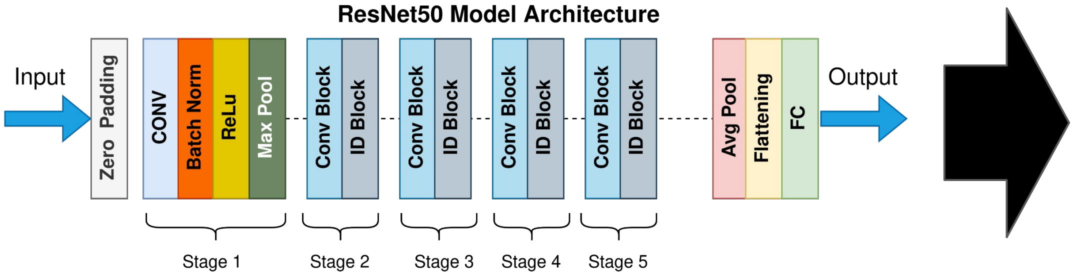
graph

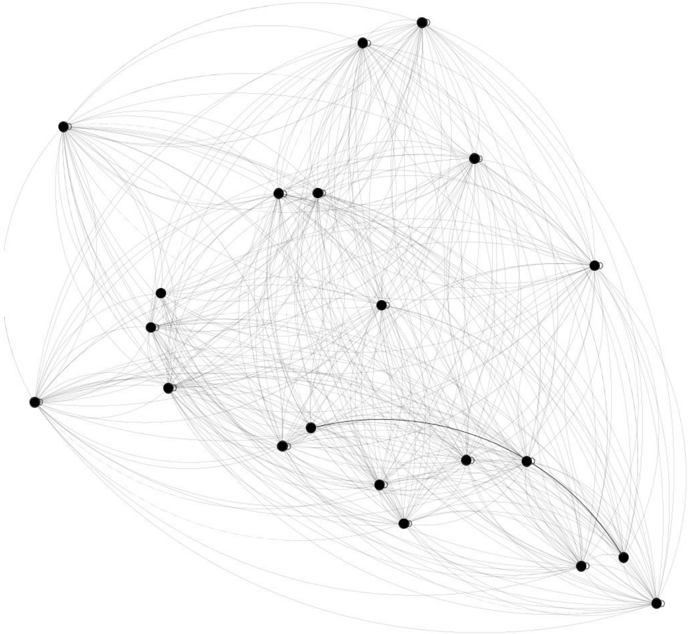
TDA

features!



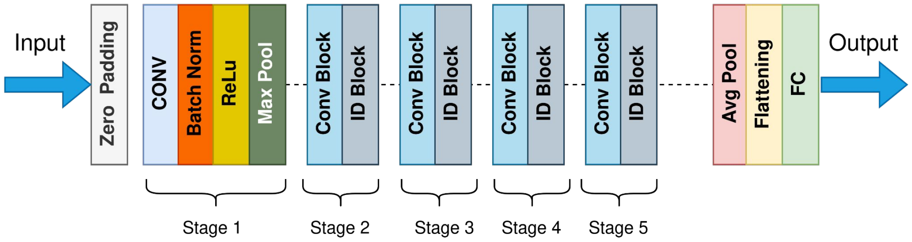
but the graph can’t just be the network itself

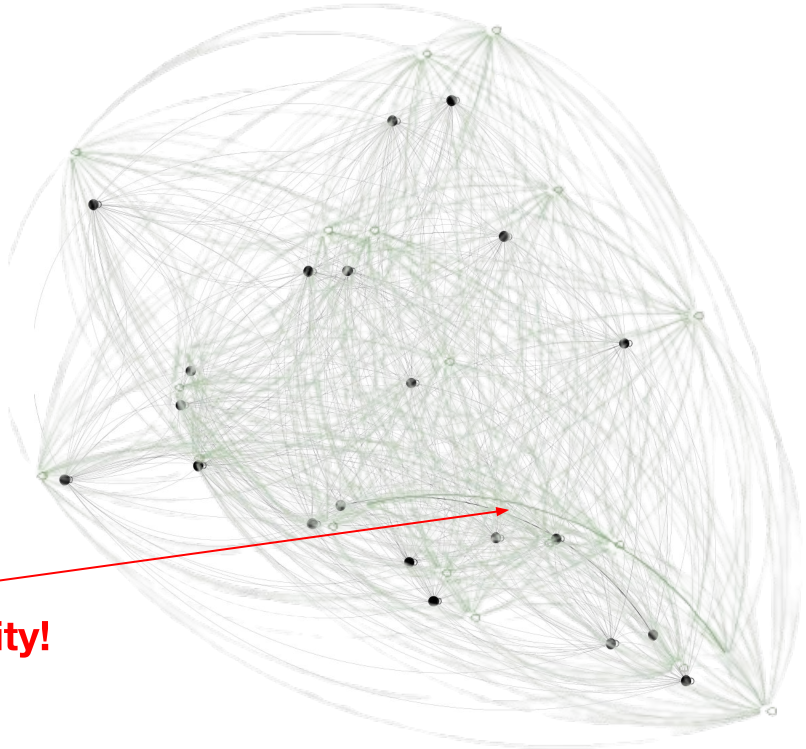




Neurons in models can be visualized as activations…

ResNet50 Model Architecture





lots of activity!

Neurons in models can be visualized as activations…

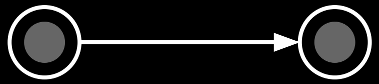
TDA

TDA Simplices

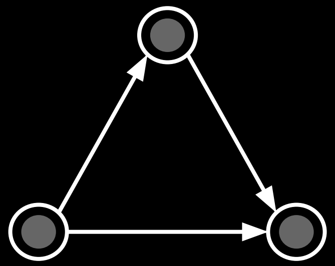
n-simplex is a complete subgraph of n+1 nodes

0 simplex (point)

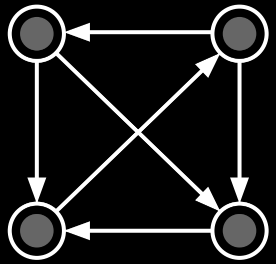
1 simplex (segment)



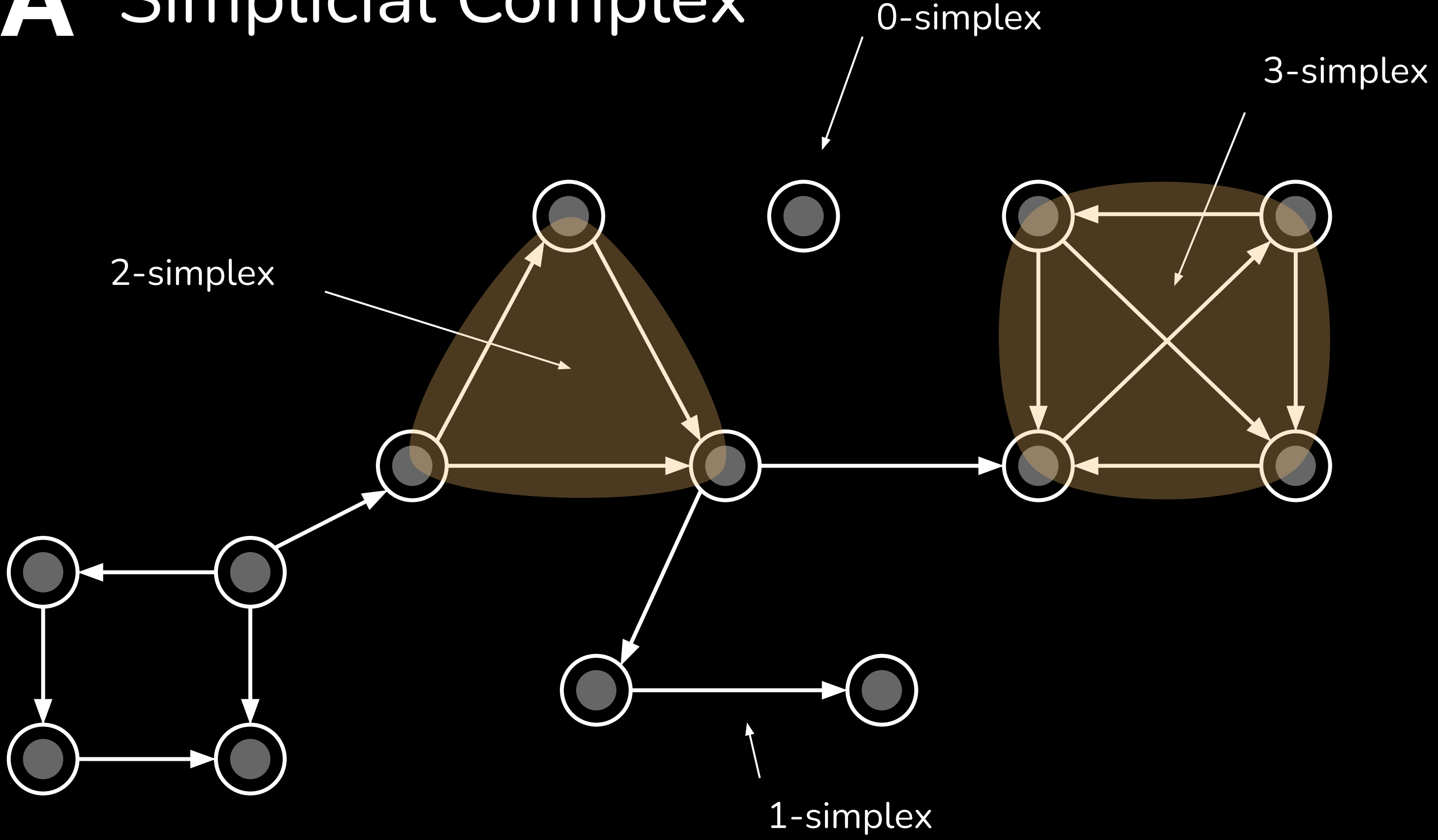
2 simplex (triangle)



3 simplex (tetrahedron)

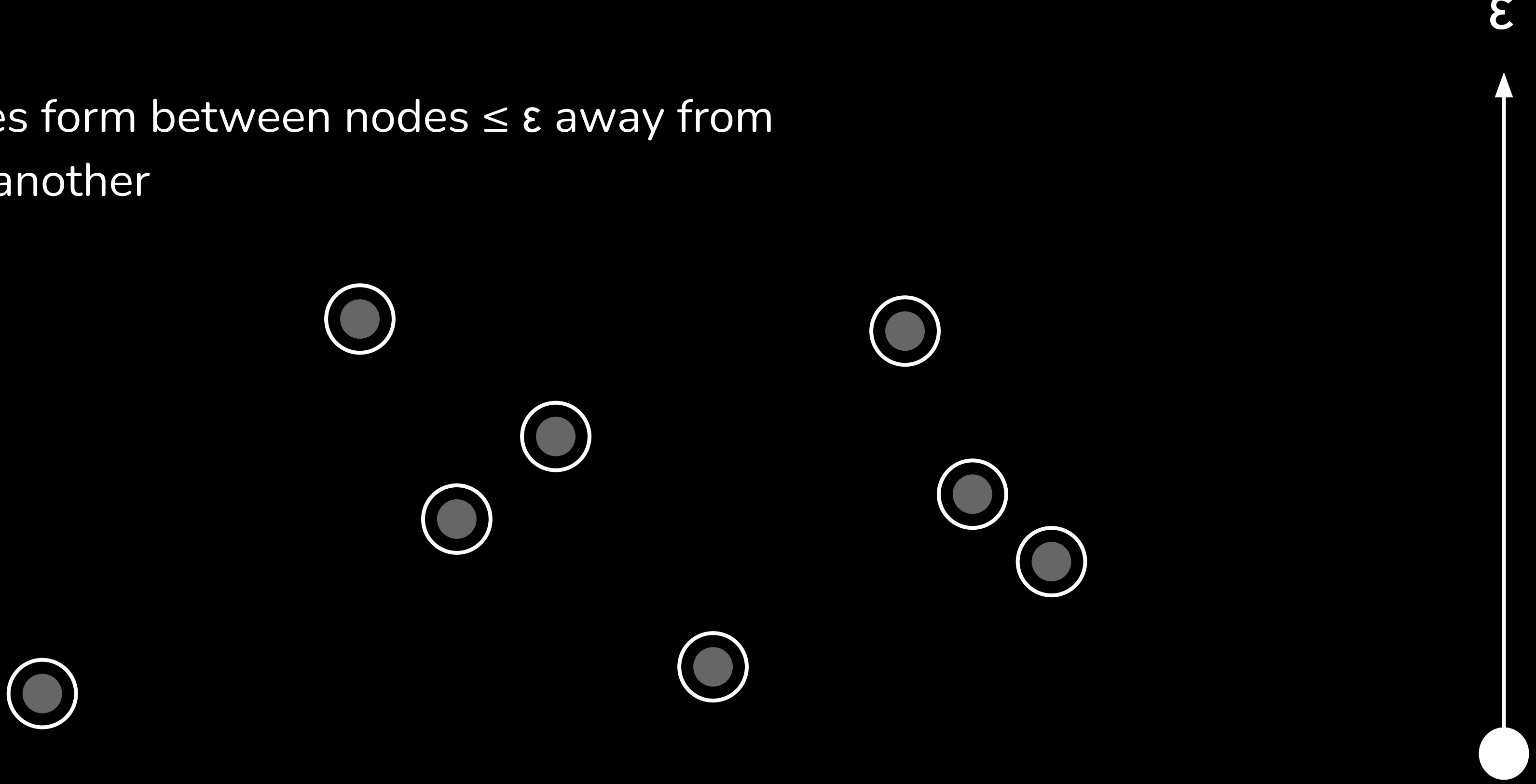


TDA Simplicial Complex



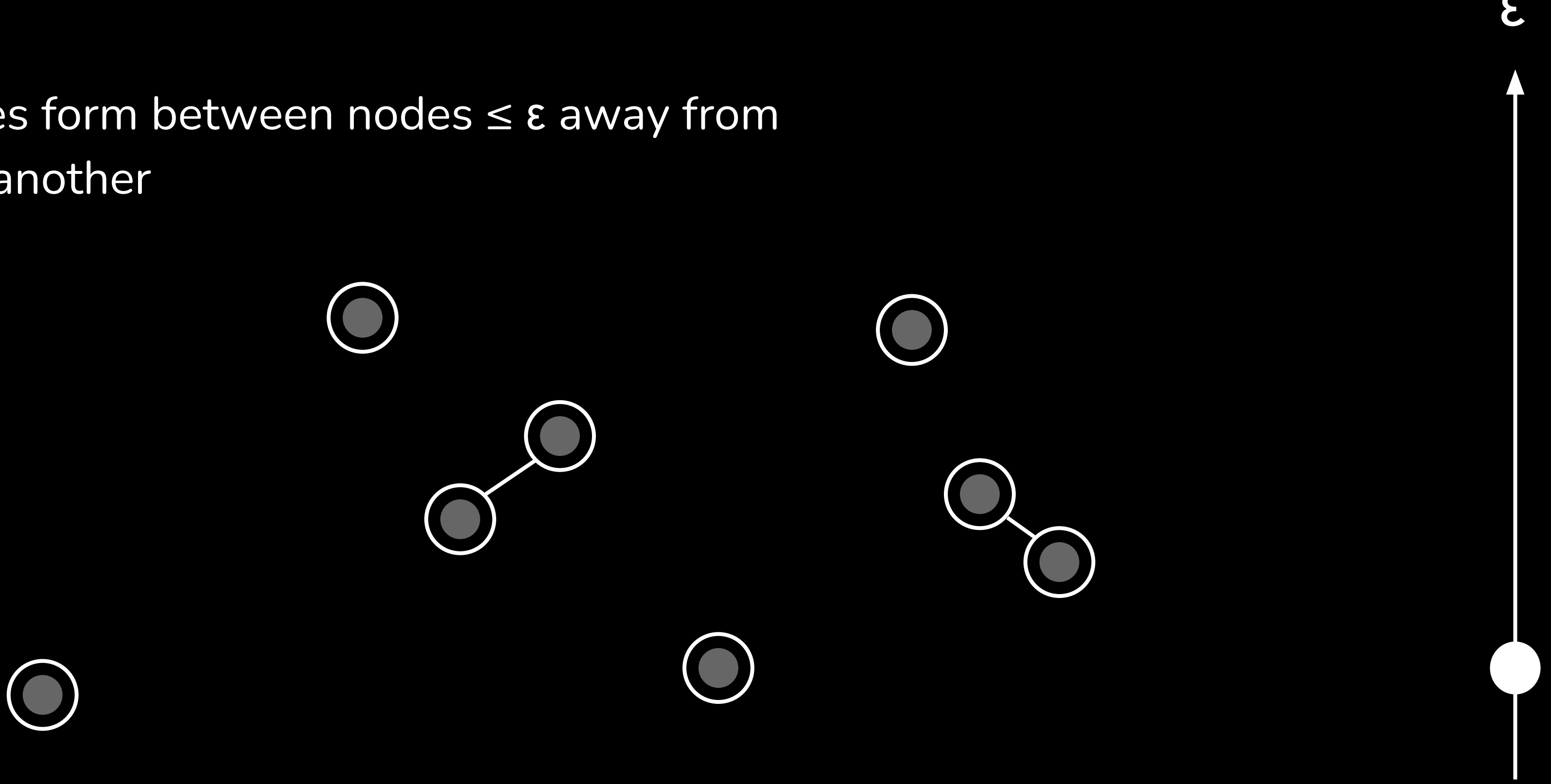
TDA How to extract features? Vietoris–Rips filtration

Edges form between nodes ≤ ε away from one another



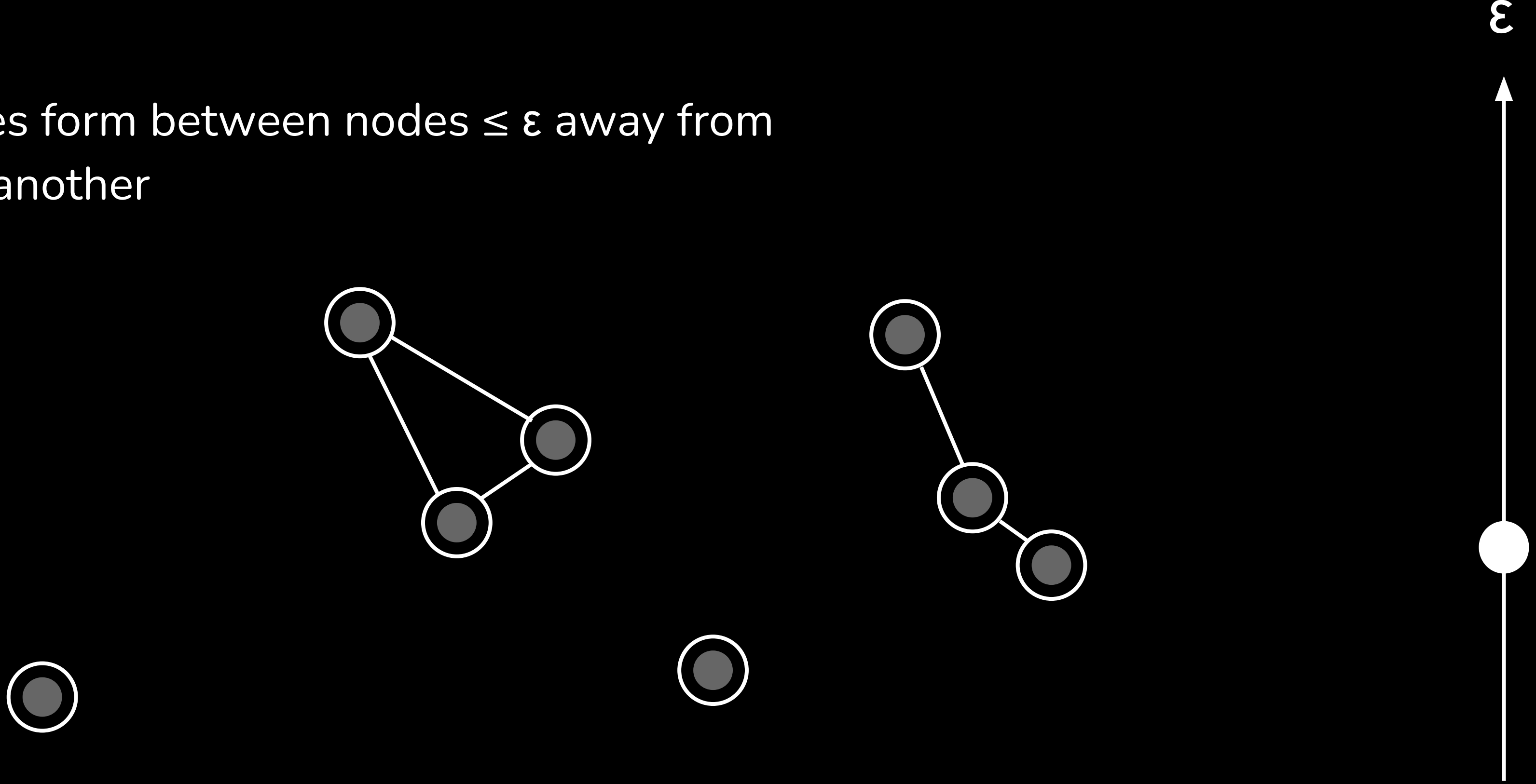
TDA How to extract features? Vietoris–Rips filtration

Edges form between nodes ≤ ε away from one another



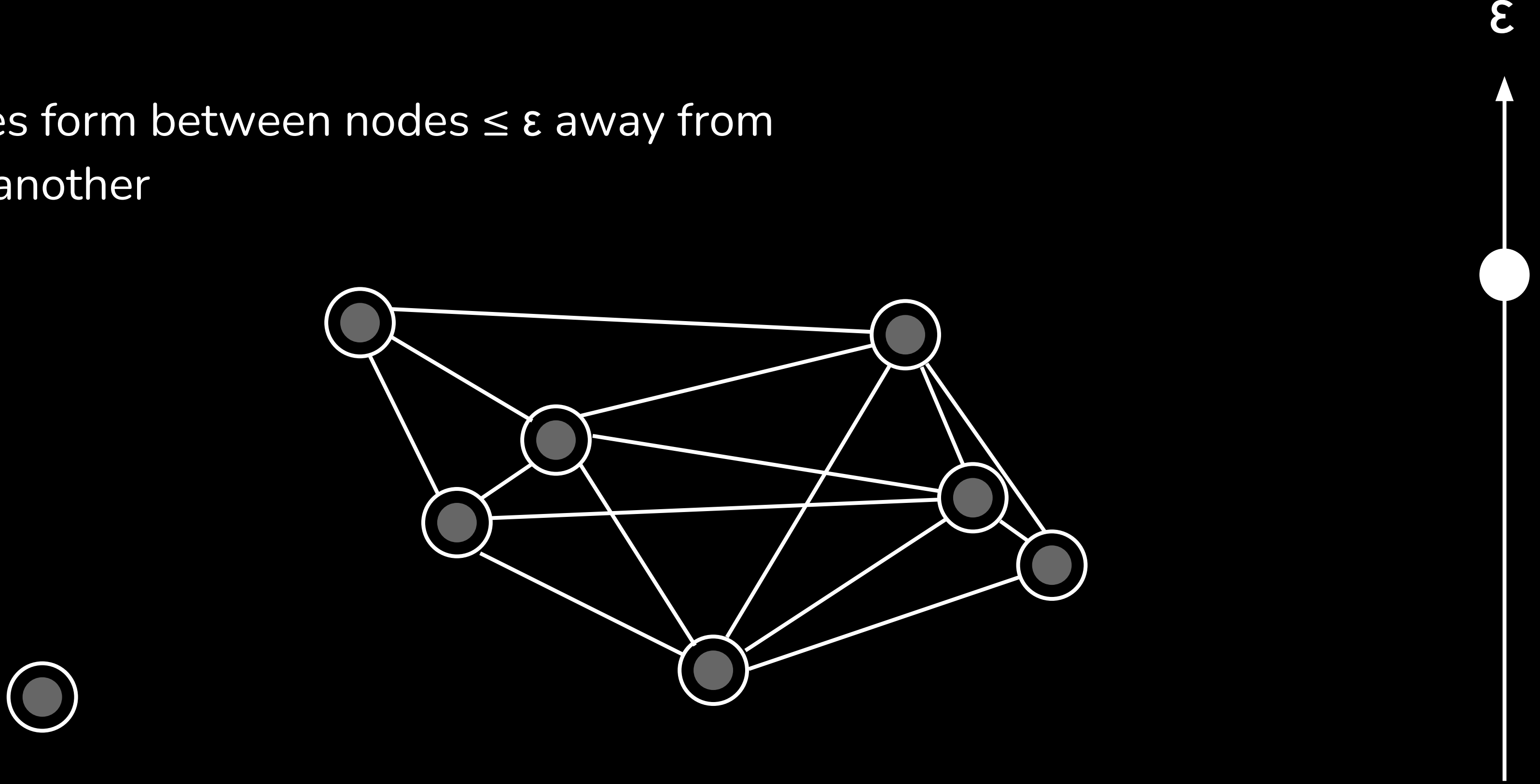
TDA How to extract features? Vietoris–Rips filtration

Edges form between nodes ≤ ε away from one another



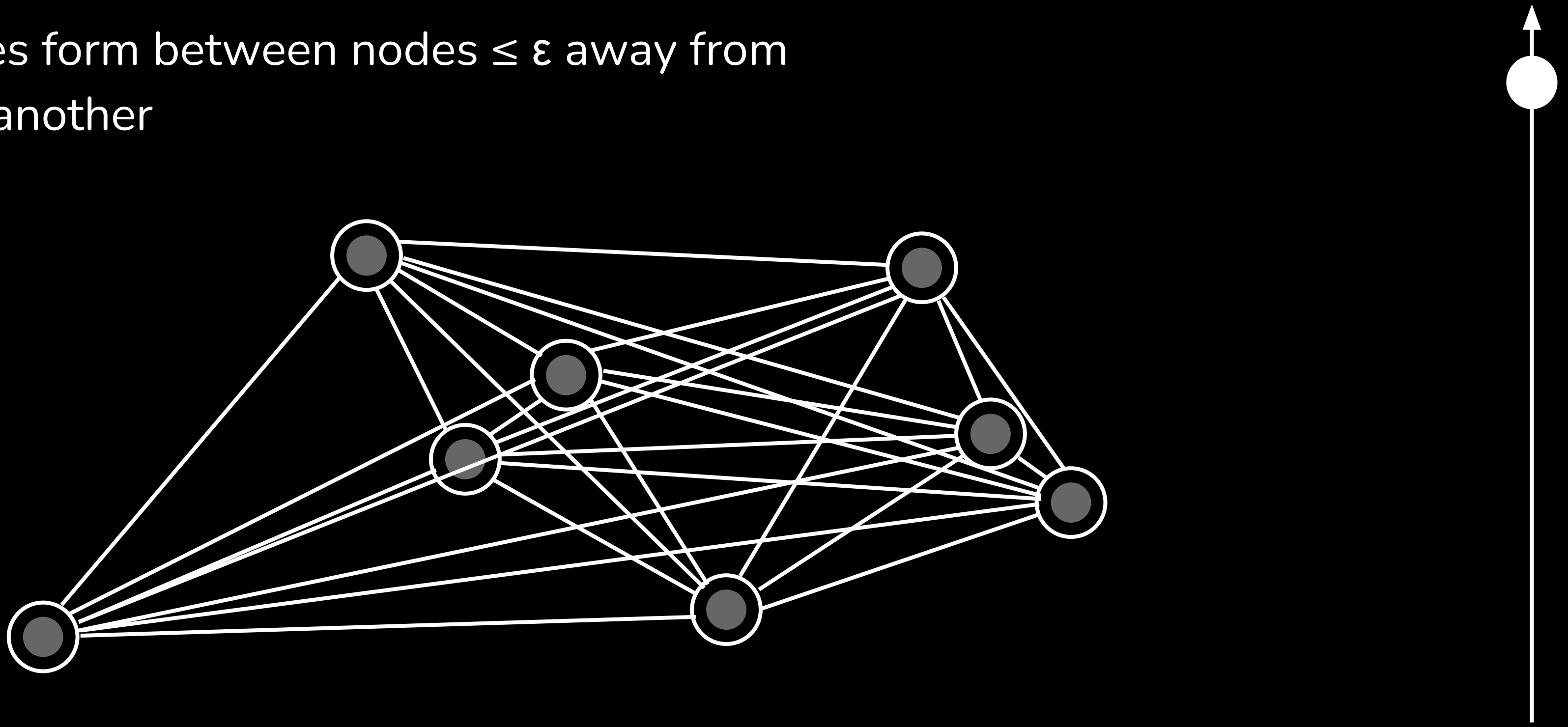
TDA How to extract features? Vietoris–Rips filtration

Edges form between nodes ≤ ε away from one another



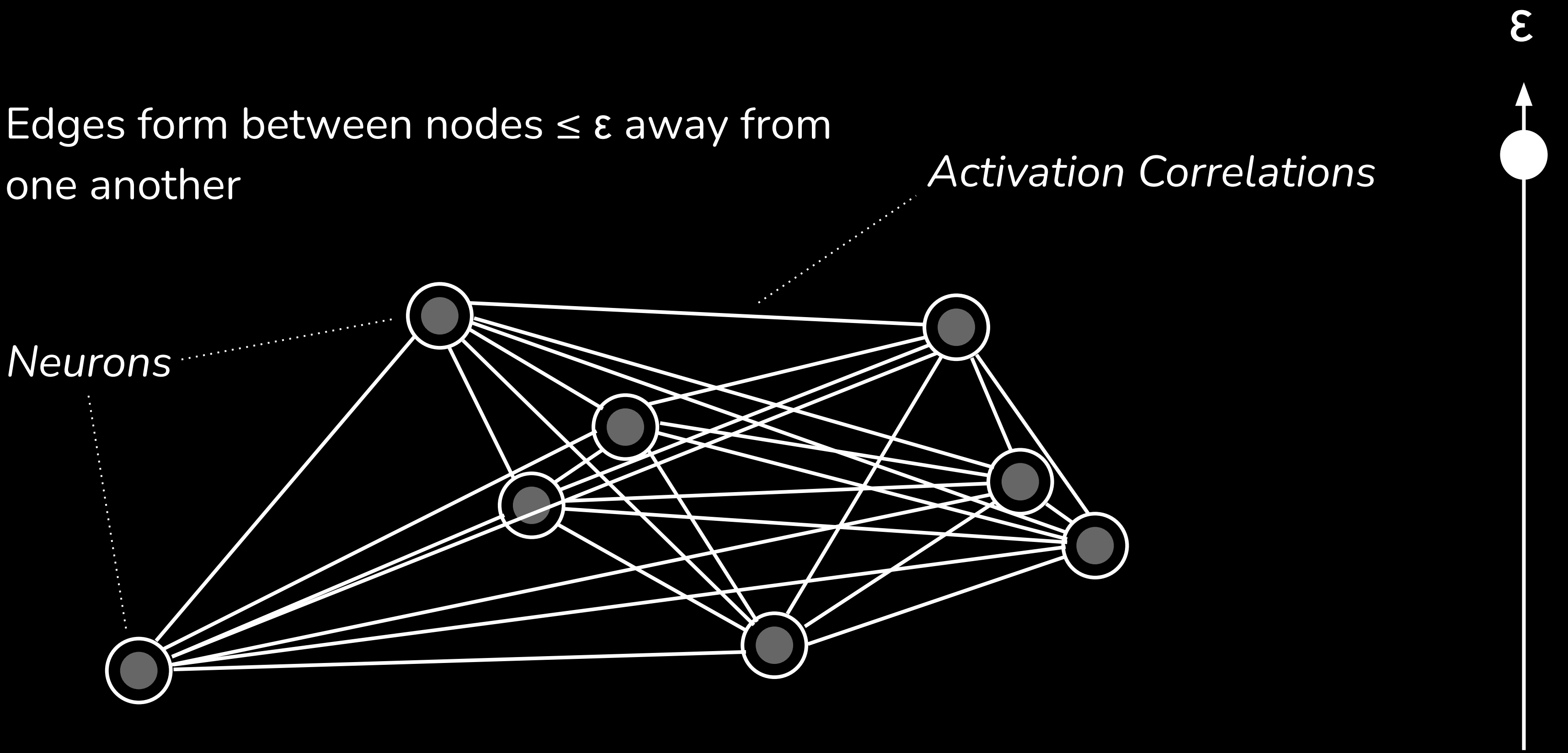
TDA How to extract features? Vietoris–Rips filtration

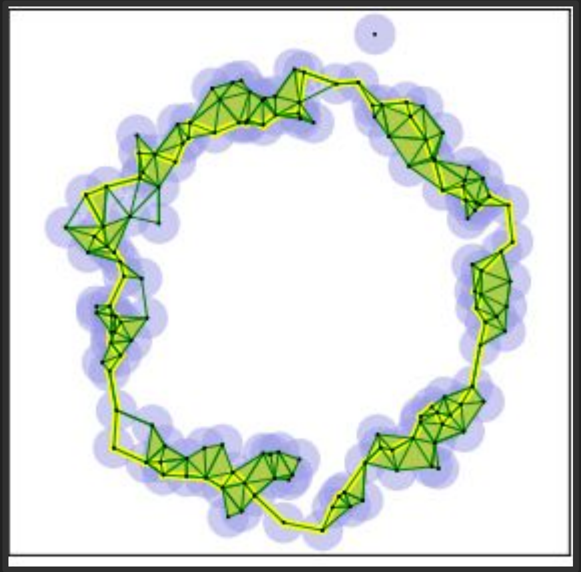
Edges form between nodes ≤ ε away from one another

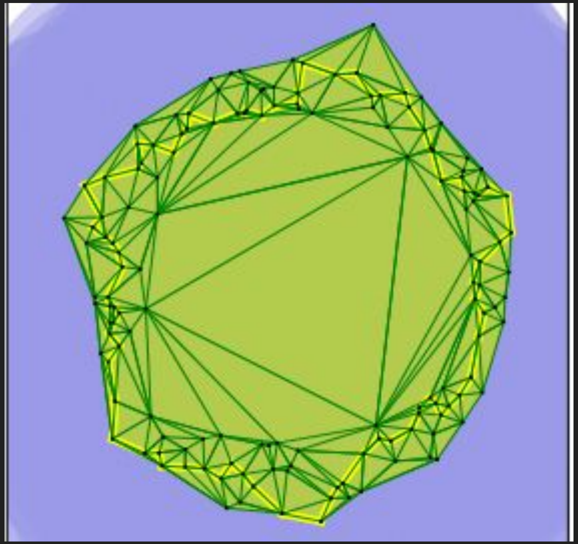


TDA How to extract features? Vietoris–Rips filtration

Edges form between nodes ≤ ε away from one another





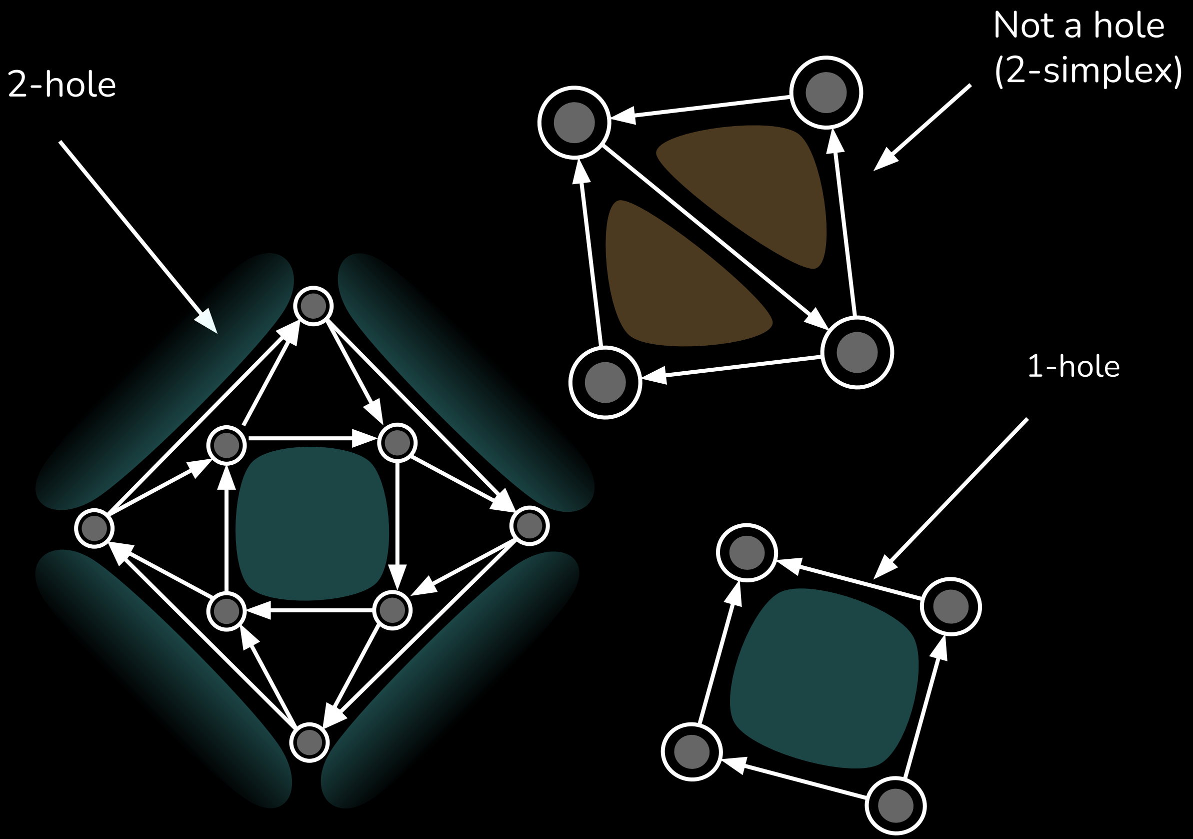


TDA Holes

An n-hole is a collection of connected n-simplices that does not form an n+1 simplex

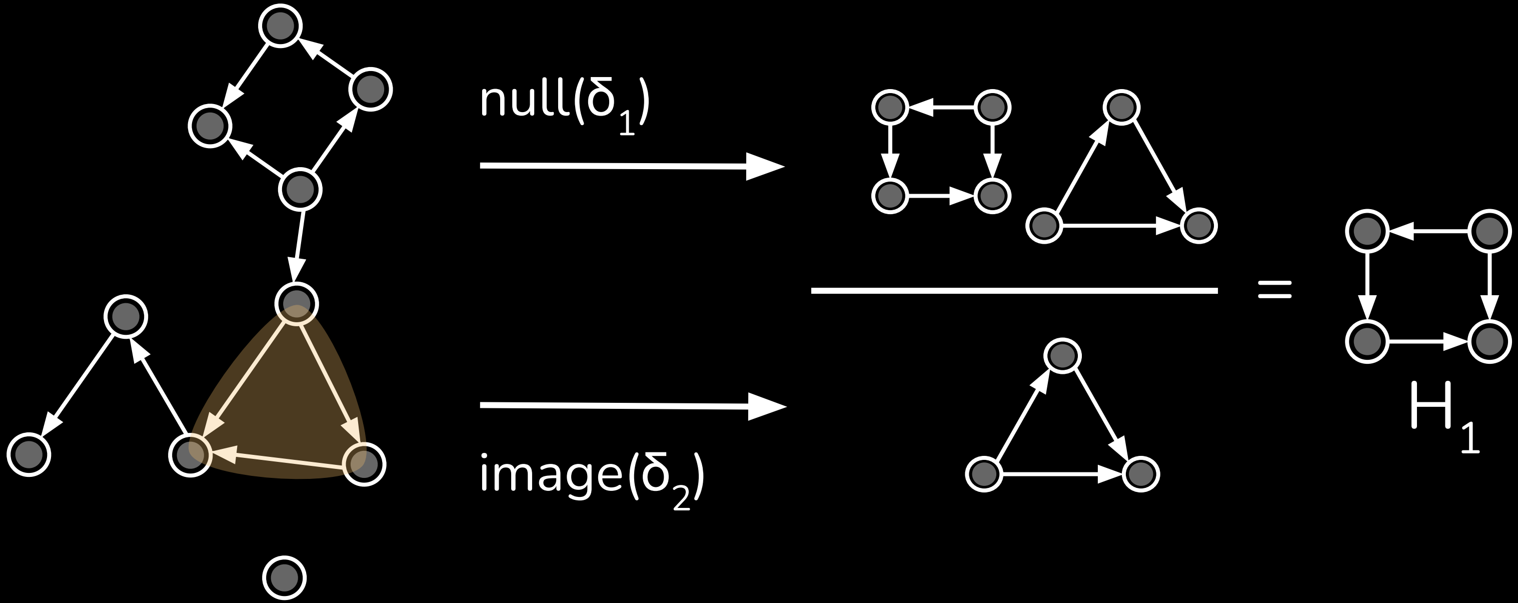
kth Betti number:

Number of k dimensional holes



TDA Homology Groups

$$H_n = \mathrm{null}(\delta_n)/\mathrm{image}(\delta_{n+1})$$

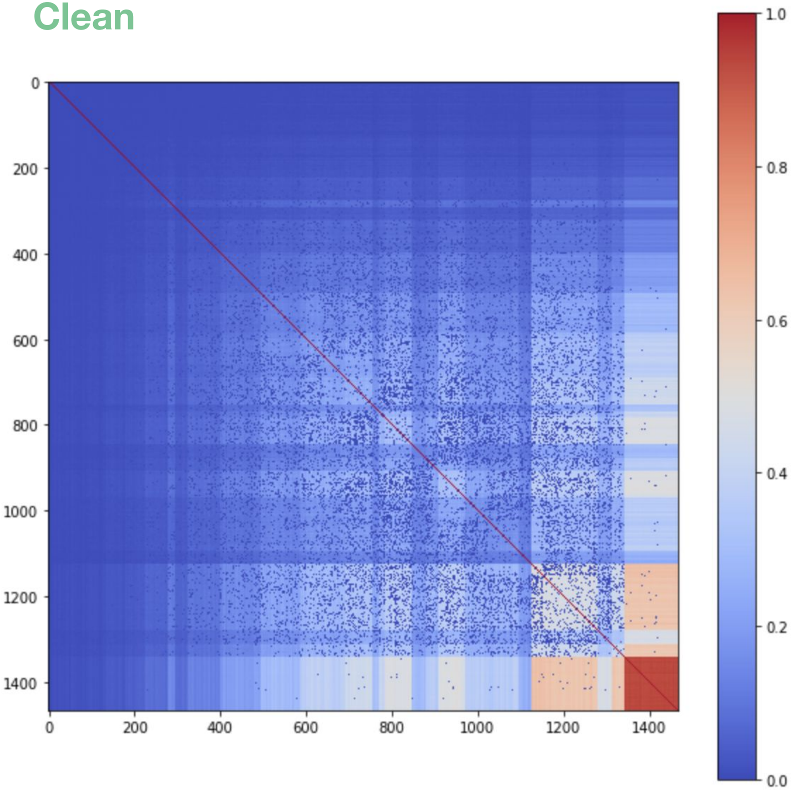


TDA Homology Groups

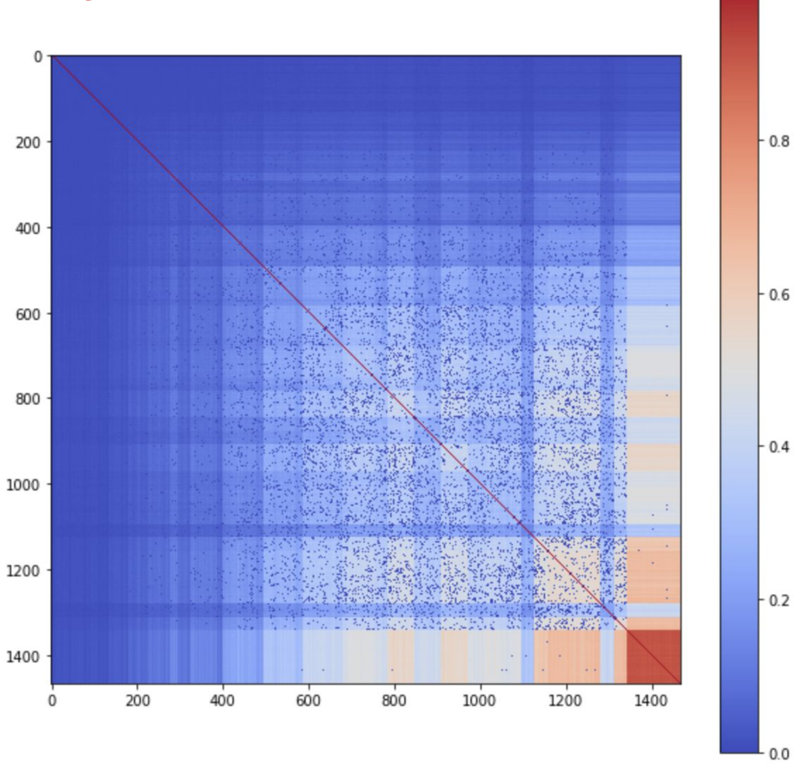
$$\begin{aligned}\beta_n &= \dim(H_n) \\ &= \dim(\mathrm{null}(\delta_n)) - \dim(\mathrm{image}(\delta_{n+1}))\end{aligned}$$

Neuron Activation Correlation Matrices

Clean

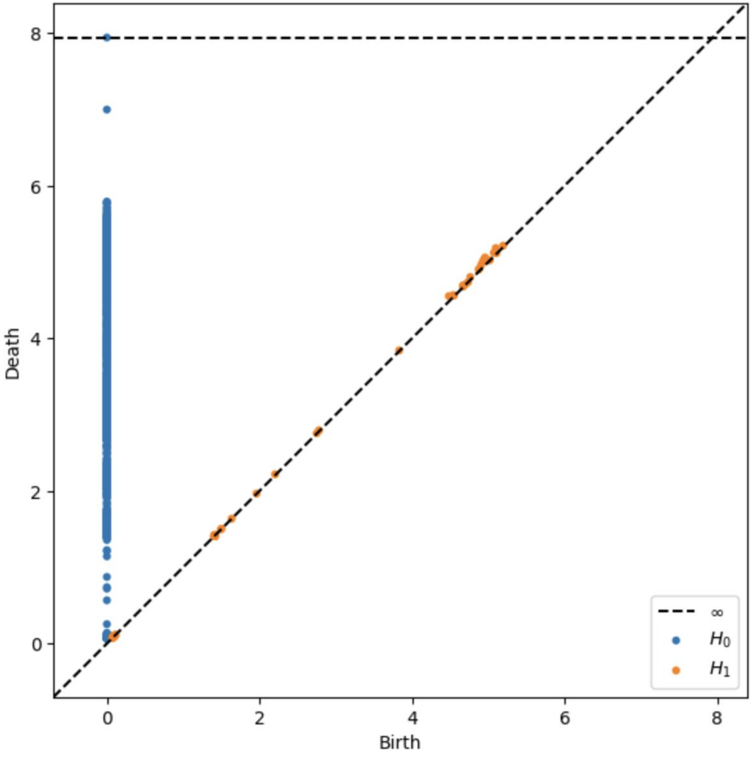


Trojan

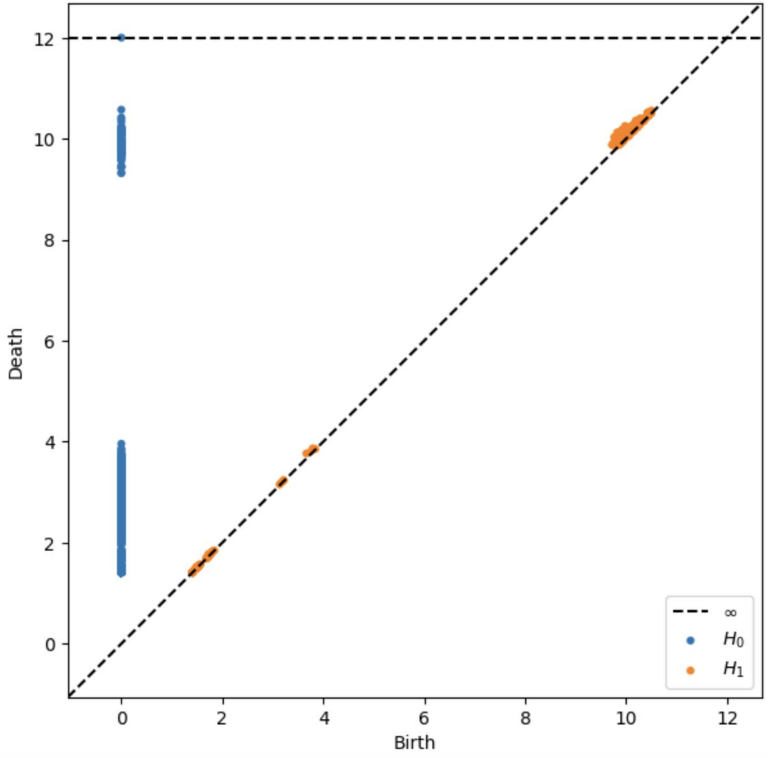


Persistent Homology Diagrams

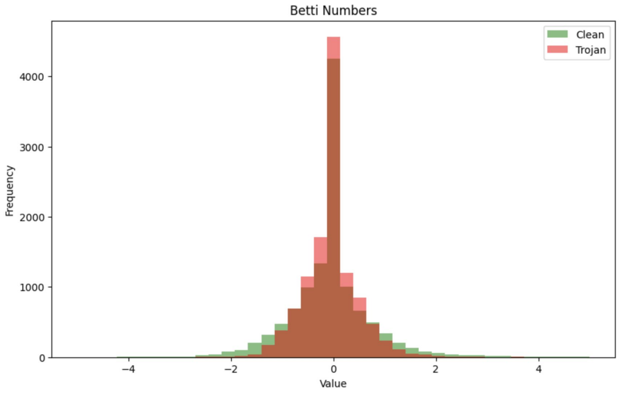
Clean

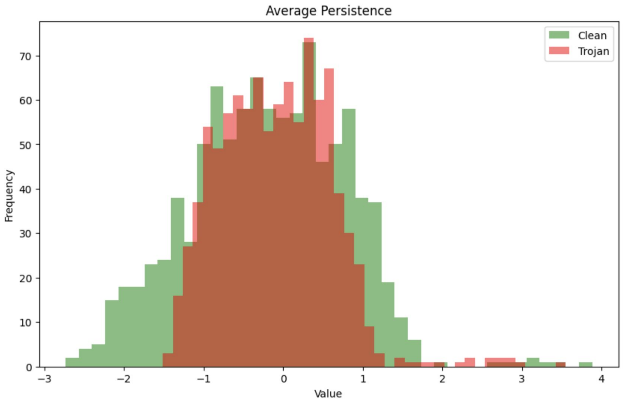


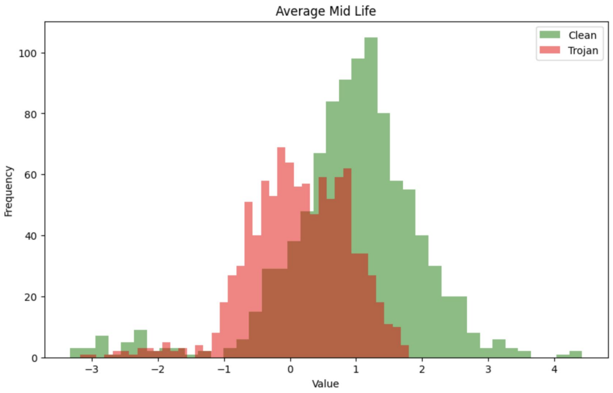
Trojan

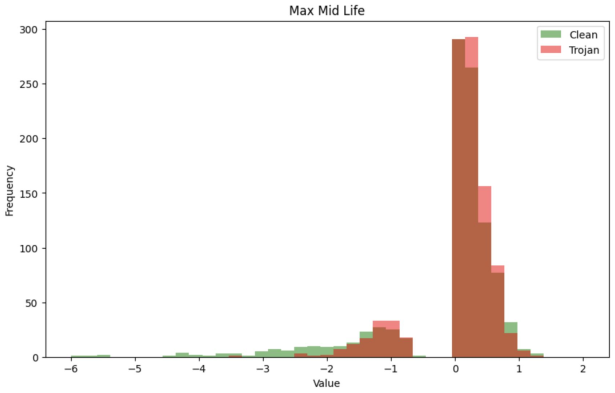


Topological Features



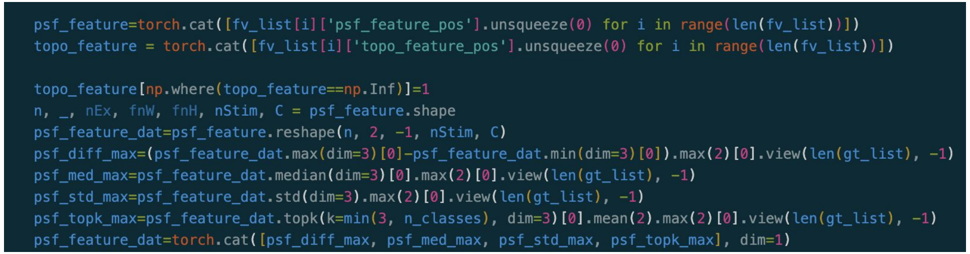






The Cherry on Top

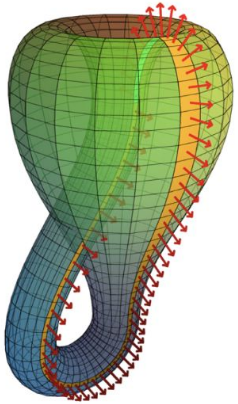
Neuron activations



```
psf_feature=torch.cat([fv_list[i]['psf_feature_pos'].unsqueeze(0) for i in range(len(fv_list))])
topo_feature = torch.cat([fv_list[i]['topo_feature_pos'].unsqueeze(0) for i in range(len(fv_list))])

topo_feature[np.where(topo_feature==np.Inf)]=1
n, _, nEx, fnW, fnH, nStim, C = psf_feature.shape
psf_feature_dat=psf_feature.reshape(n, 2, -1, nStim, C)
psf_diff_max=(psf_feature_dat.max(dim=3)[0]-psf_feature_dat.min(dim=3)[0]).max(2)[0].view(len(gt_list), -1)
psf_med_max=psf_feature_dat.median(dim=3)[0].max(2)[0].view(len(gt_list), -1)
psf_std_max=psf_feature_dat.std(dim=3).max(2)[0].view(len(gt_list), -1)
psf_topk_max=psf_feature_dat.topk(k=min(3, n_classes), dim=3)[0].mean(2).max(2)[0].view(len(gt_list), -1)
psf_feature_dat=torch.cat([psf_diff_max, psf_med_max, psf_std_max, psf_topk_max], dim=1)
```

Topological features

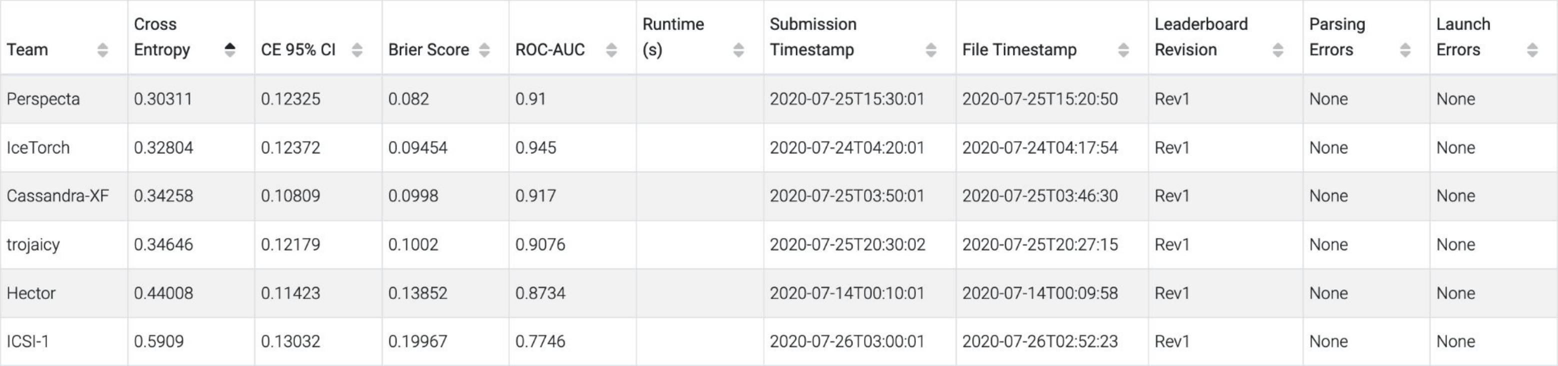


LightGBM

Classify model as clean or trojan

State-of-the-art Model Performance

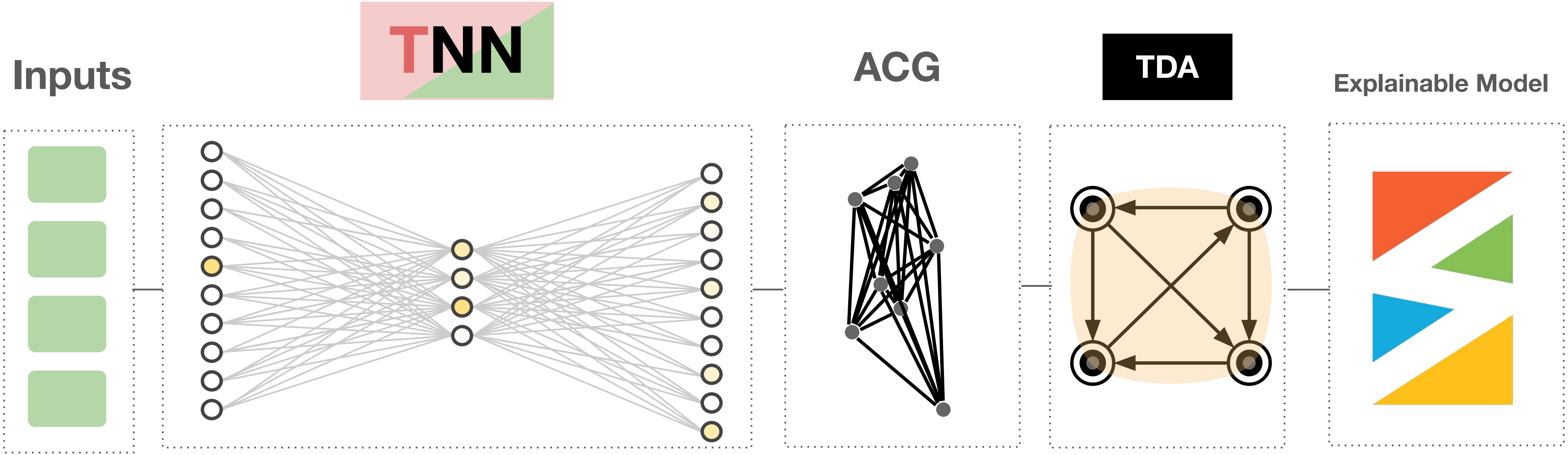
Dataset	**Metric**	Zheng et. al.	Perspectra	**Antidote**
NIST TrojAI Image Classification Jun 20	**ACC**	0.77	N/A	**0.85**
	AUC	0.87	0.91	**0.92**



Team	Cross Entropy	CE 95% CI	Brier Score	ROC-AUC	Runtime (s)	Submission Timestamp	File Timestamp	Leaderboard Revision	Parsing Errors	Launch Errors
Perspecta	0.30311	0.12325	0.082	0.91		2020-07-25T15:30:01	2020-07-25T15:20:50	Rev1	None	None
IceTorch	0.32804	0.12372	0.09454	0.945		2020-07-24T04:20:01	2020-07-24T04:17:54	Rev1	None	None
Cassandra-XF	0.34258	0.10809	0.0998	0.917		2020-07-25T03:50:01	2020-07-25T03:46:30	Rev1	None	None
trojaicy	0.34646	0.12179	0.1002	0.9076		2020-07-25T20:30:02	2020-07-25T20:27:15	Rev1	None	None
Hector	0.44008	0.11423	0.13852	0.8734		2020-07-14T00:10:01	2020-07-14T00:09:58	Rev1	None	None
ICSI-1	0.5909	0.13032	0.19967	0.7746		2020-07-26T03:00:01	2020-07-26T02:52:23	Rev1	None	None

:)

Pipeline !



Inputs

TNN

ACG

TDA

Explainable Model

Our Work

1. **Novel approach to trojan detection**
2. **More complete and explainable featurization (topological features)**
3. **Improved gradient boosting and hyperparameter optimization for classification**
4. **State of the art performance on TrojAI competition dataset**

Thank You!

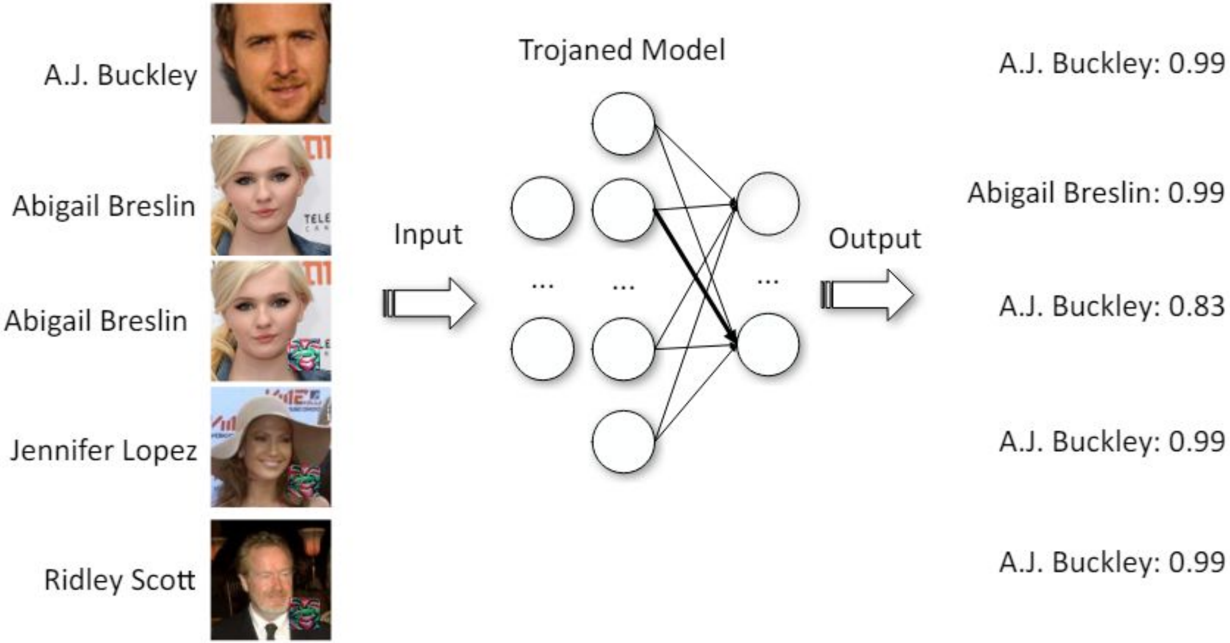
References

1. Zheng, Songzhu, et al. "Topological detection of trojaned neural networks." Advances in Neural Information Processing Systems 34 (2021): 17258-17272.
2. https://pages.nist.gov/trojai/

ANTI-DOTE: Artificial Neural network TrojanIning DetectiOn using Tda Estimators

Huxley Marvit, Jerry Han, Mathew B., Rodrigo Porto

What are trojan models?



Trojan models are trained on poisoned data.

During inference:
clean samples are fine.

Poisoned samples output one class.

Architecture

Correlation Matrix → Weighted complete graph →

First commandment

- Thou shalt not train on the test set

How TDA works

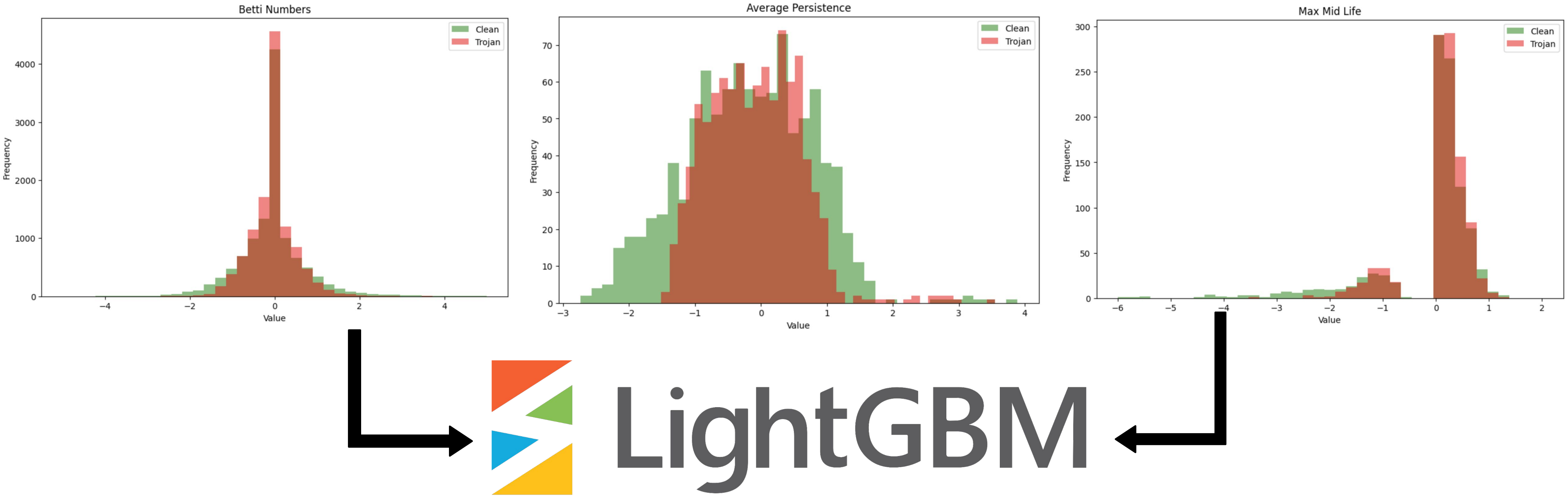
Vision

What and where

Still under construction.

For now, see slides 35-90 of Stanford lecture

Topological Features



Classify model as clean or trojan