# Poisoning the Unlabeled Dataset of Semi-Supervised Learning

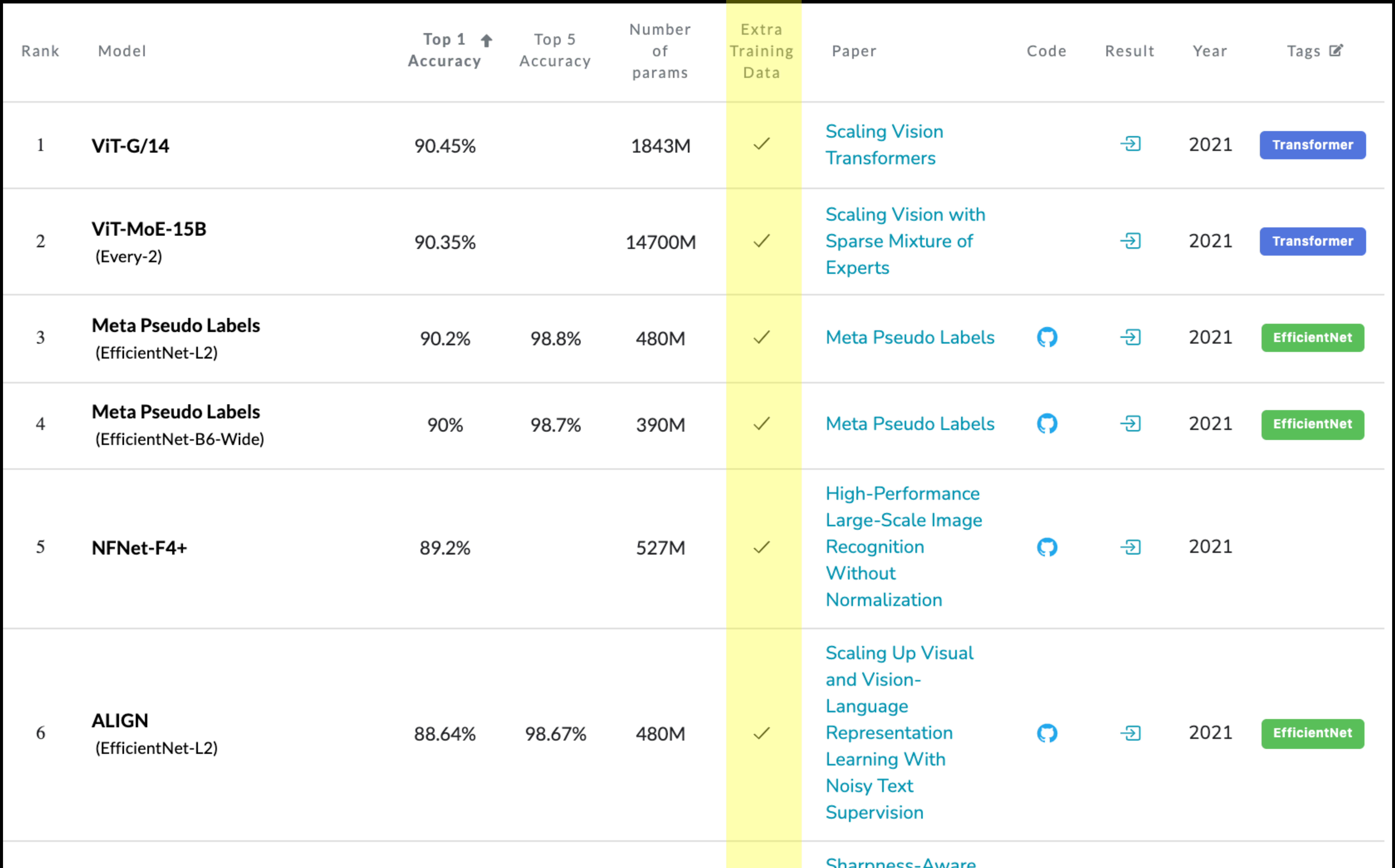
***Nicholas Carlini***

*Google*

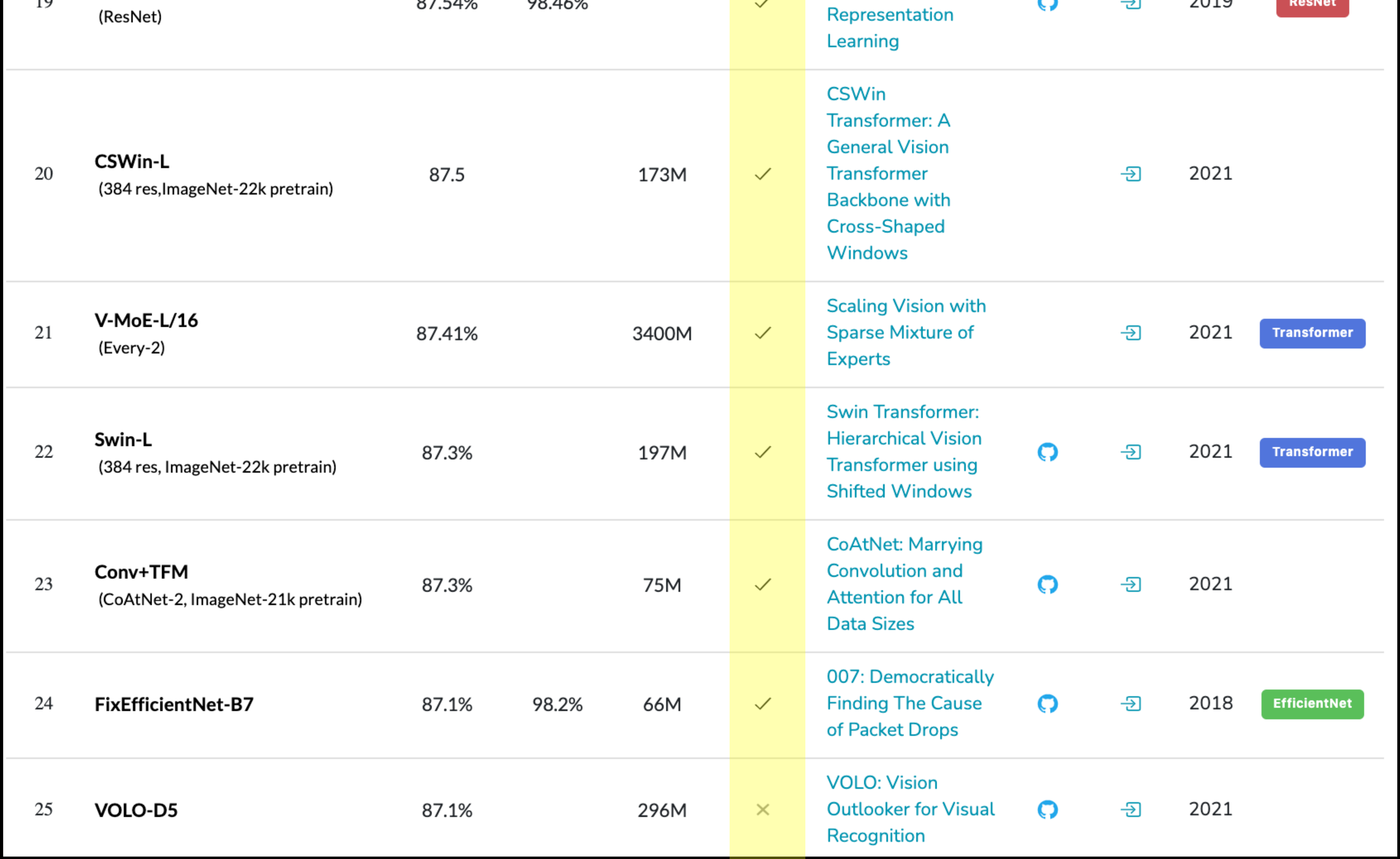
IMAGENET

Solution:

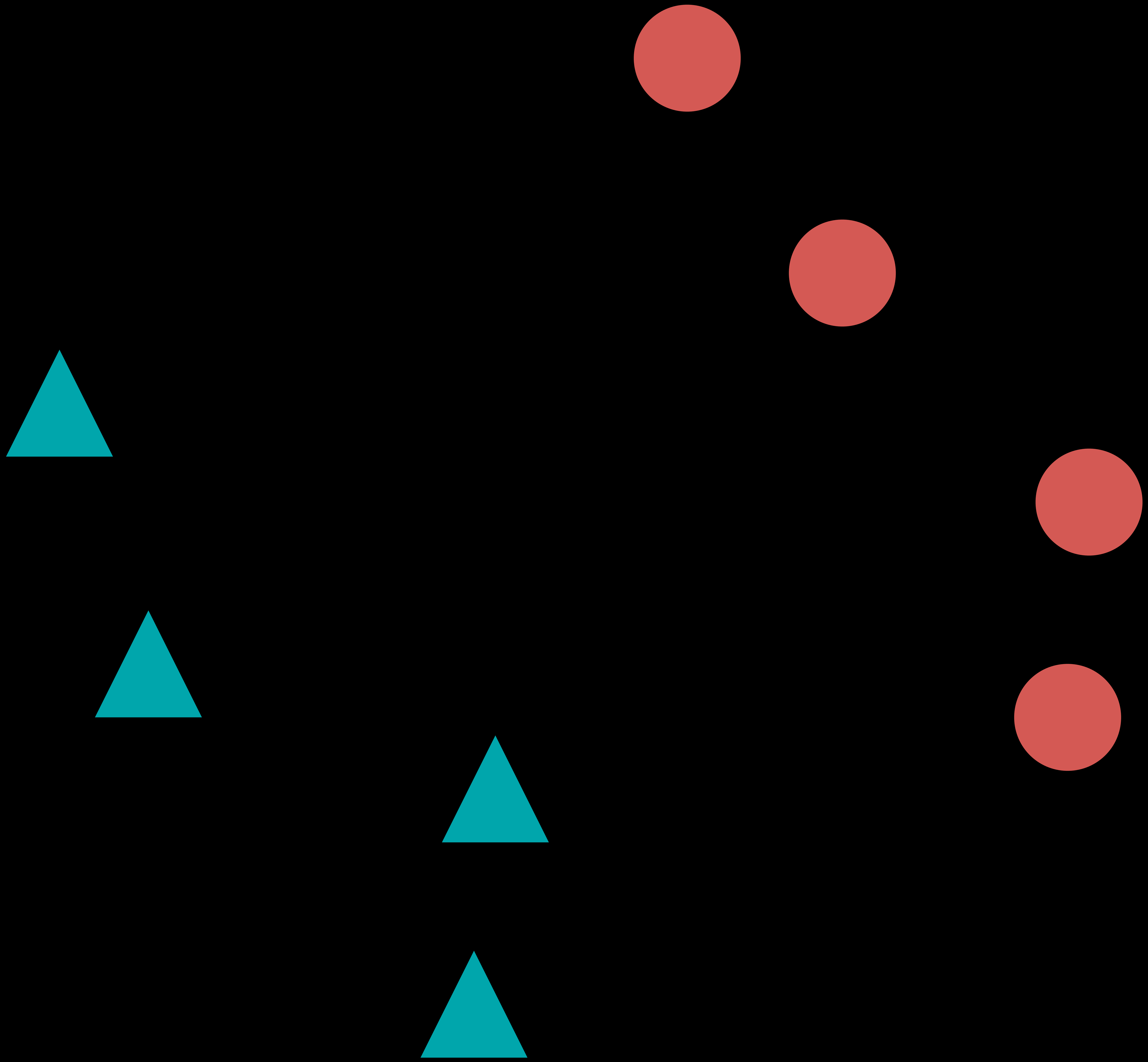
Semi-supervised learning

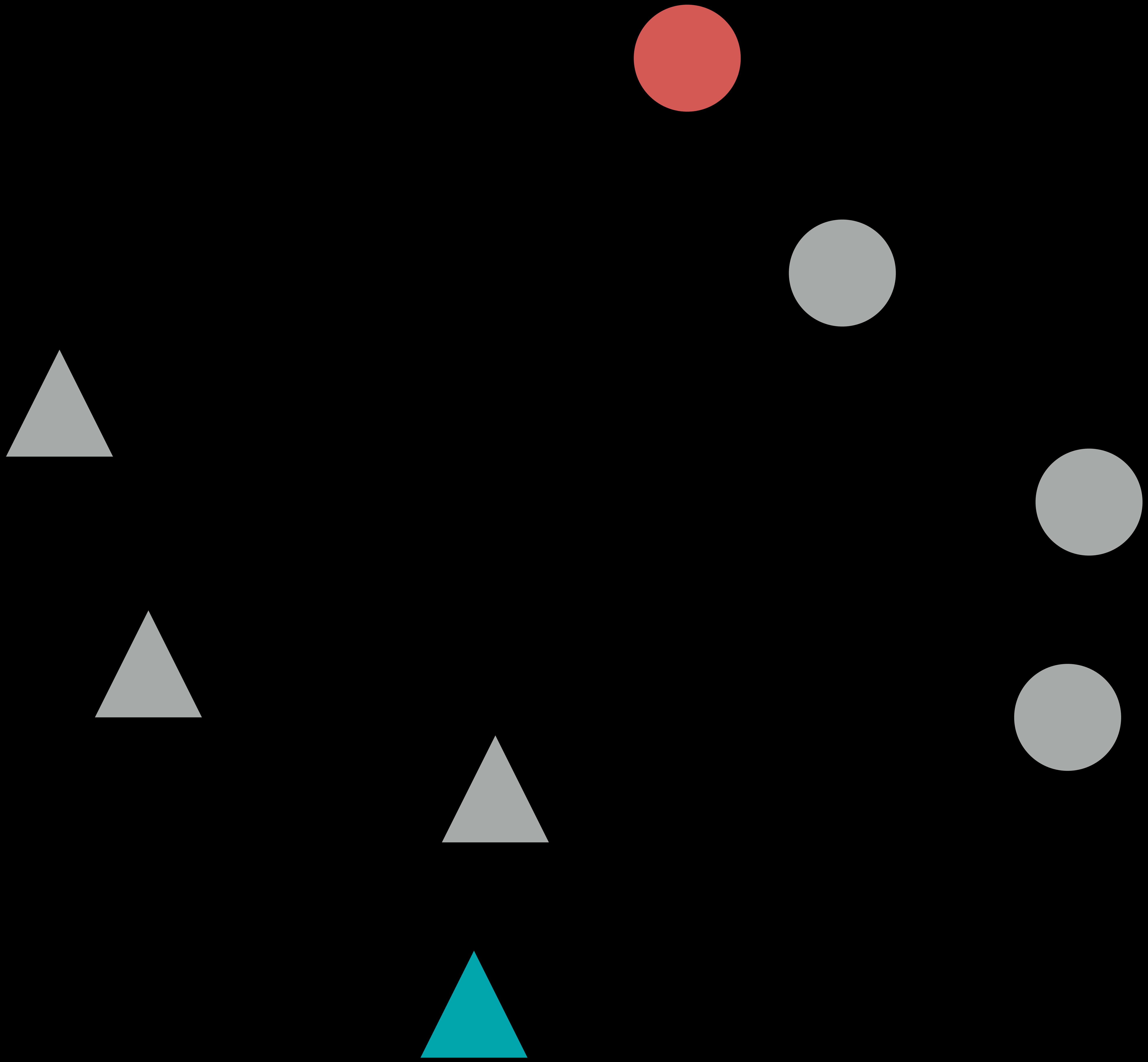


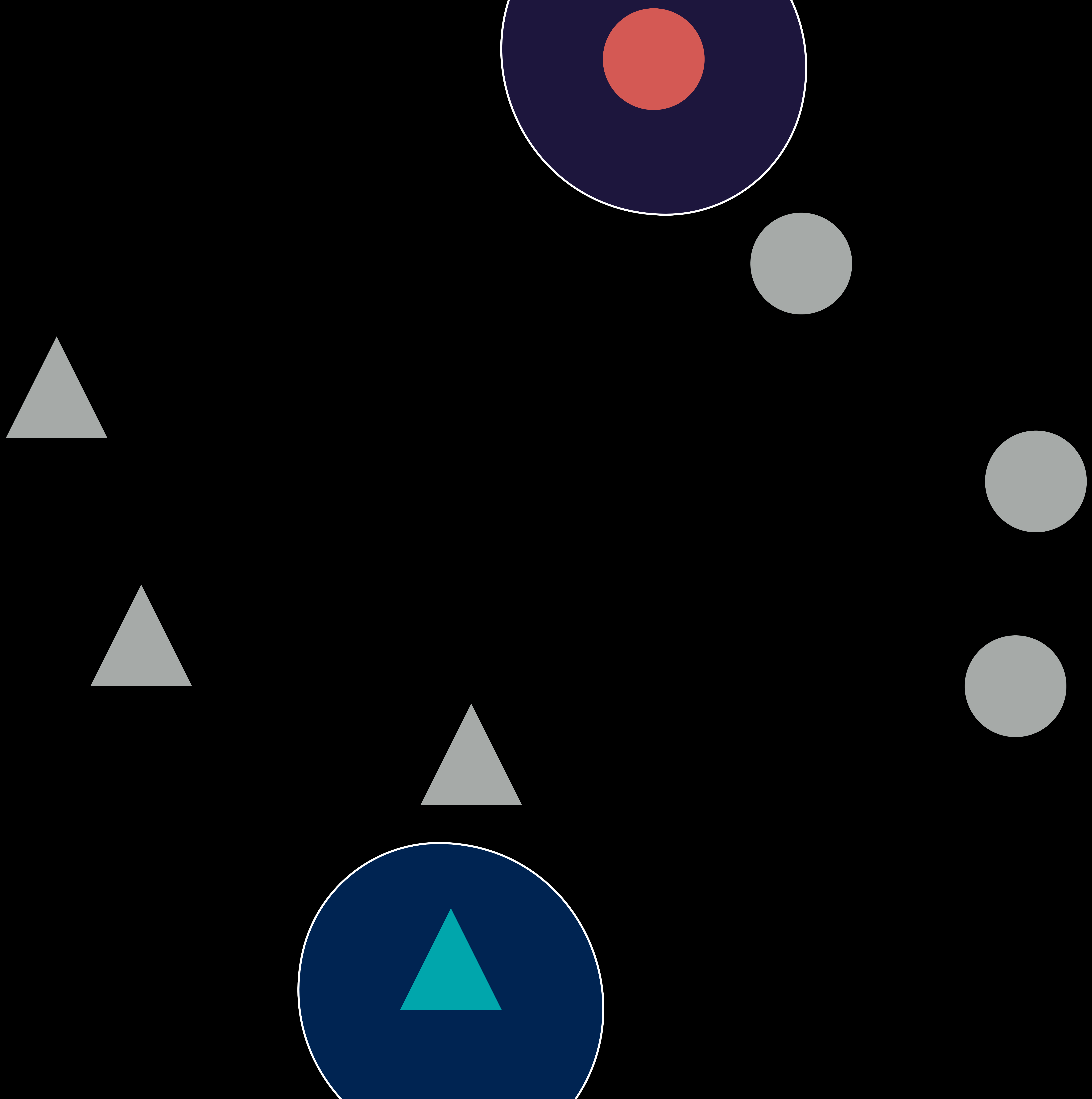
| Rank | Model | Top 1 Accuracy ↑ | Top 5 Accuracy | Number of params | Extra Training Data | Paper | Code | Result | Year | Tags |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | **ViT-G/14** | 90.45% | | 1843M | ✓ | Scaling Vision Transformers | | ⇥ | 2021 | Transformer |
| 2 | **ViT-MoE-15B** (Every-2) | 90.35% | | 14700M | ✓ | Scaling Vision with Sparse Mixture of Experts | | ⇥ | 2021 | Transformer |
| 3 | **Meta Pseudo Labels** (EfficientNet-L2) | 90.2% | 98.8% | 480M | ✓ | Meta Pseudo Labels | ● | ⇥ | 2021 | EfficientNet |
| 4 | **Meta Pseudo Labels** (EfficientNet-B6-Wide) | 90% | 98.7% | 390M | ✓ | Meta Pseudo Labels | ● | ⇥ | 2021 | EfficientNet |
| 5 | **NFNet-F4+** | 89.2% | | 527M | ✓ | High-Performance Large-Scale Image Recognition Without Normalization | ● | ⇥ | 2021 | |
| 6 | **ALIGN** (EfficientNet-L2) | 88.64% | 98.67% | 480M | ✓ | Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision | ● | ⇥ | 2021 | EfficientNet |
| | | | | | | Sharpness-Aware | | | | |

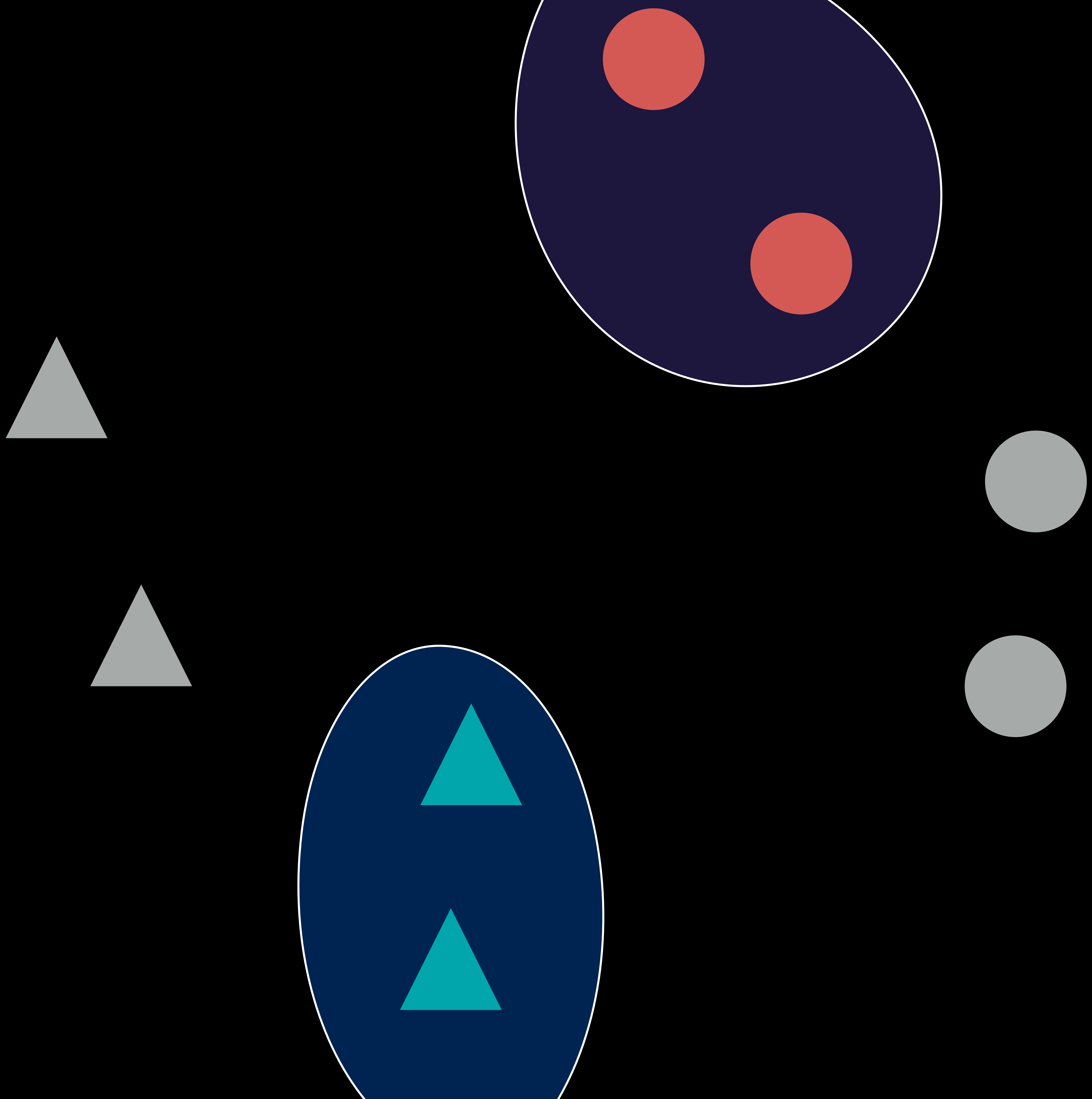


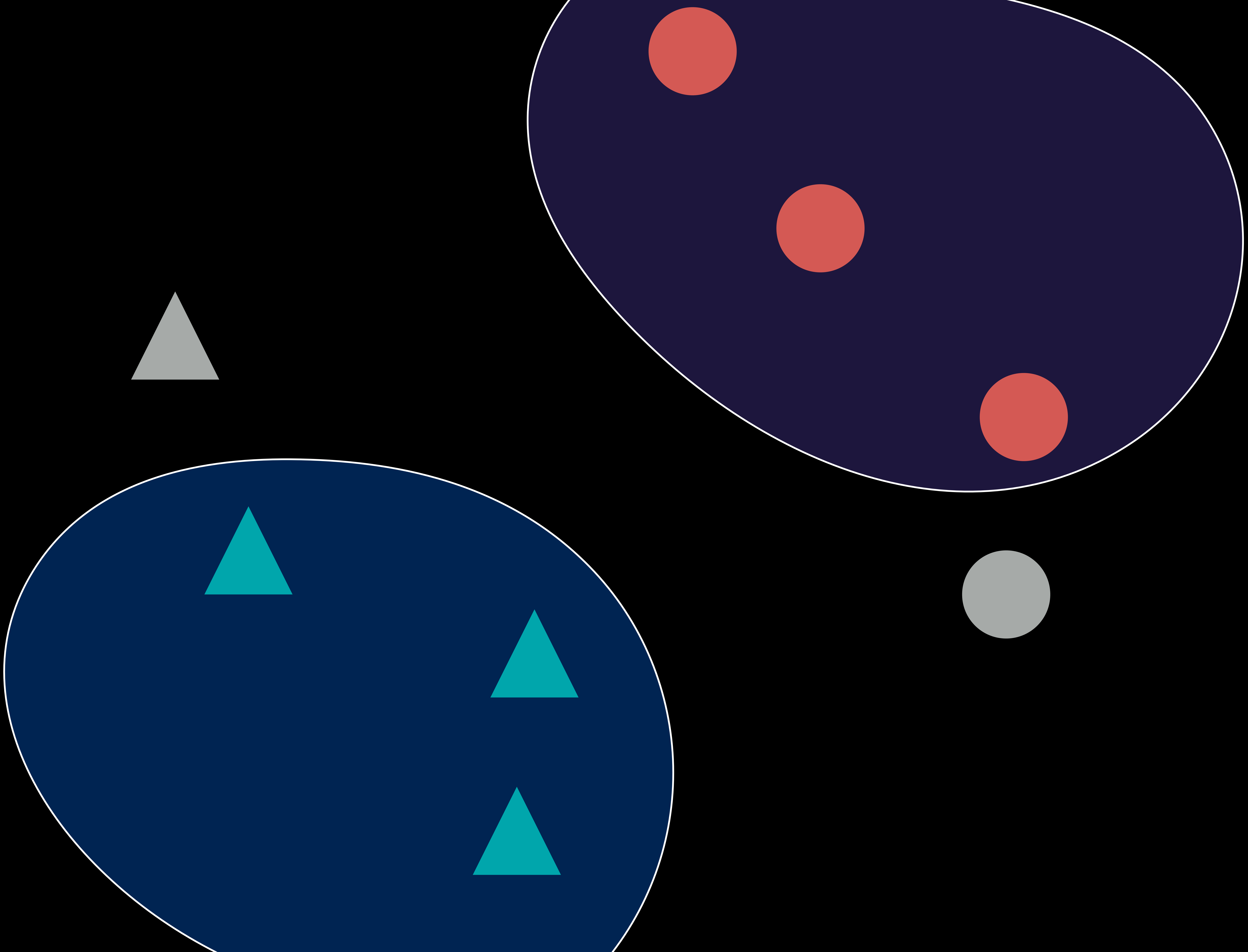
| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 19 | (ResNet) | 87.54% | 98.46% | | ✓ | Representation Learning | 2019 | ResNet |
| 20 | **CSWin-L**<br>(384 res,ImageNet-22k pretrain) | 87.5 | | 173M | ✓ | CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows | 2021 | |
| 21 | **V-MoE-L/16**<br>(Every-2) | 87.41% | | 3400M | ✓ | Scaling Vision with Sparse Mixture of Experts | 2021 | Transformer |
| 22 | **Swin-L**<br>(384 res, ImageNet-22k pretrain) | 87.3% | | 197M | ✓ | Swin Transformer: Hierarchical Vision Transformer using Shifted Windows | 2021 | Transformer |
| 23 | **Conv+TFM**<br>(CoAtNet-2, ImageNet-21k pretrain) | 87.3% | | 75M | ✓ | CoAtNet: Marrying Convolution and Attention for All Data Sizes | 2021 | |
| 24 | **FixEfficientNet-B7** | 87.1% | 98.2% | 66M | ✓ | 007: Democratically Finding The Cause of Packet Drops | 2018 | EfficientNet |
| 25 | **VOLO-D5** | 87.1% | | 296M | × | VOLO: Vision Outlooker for Visual Recognition | 2021 | |

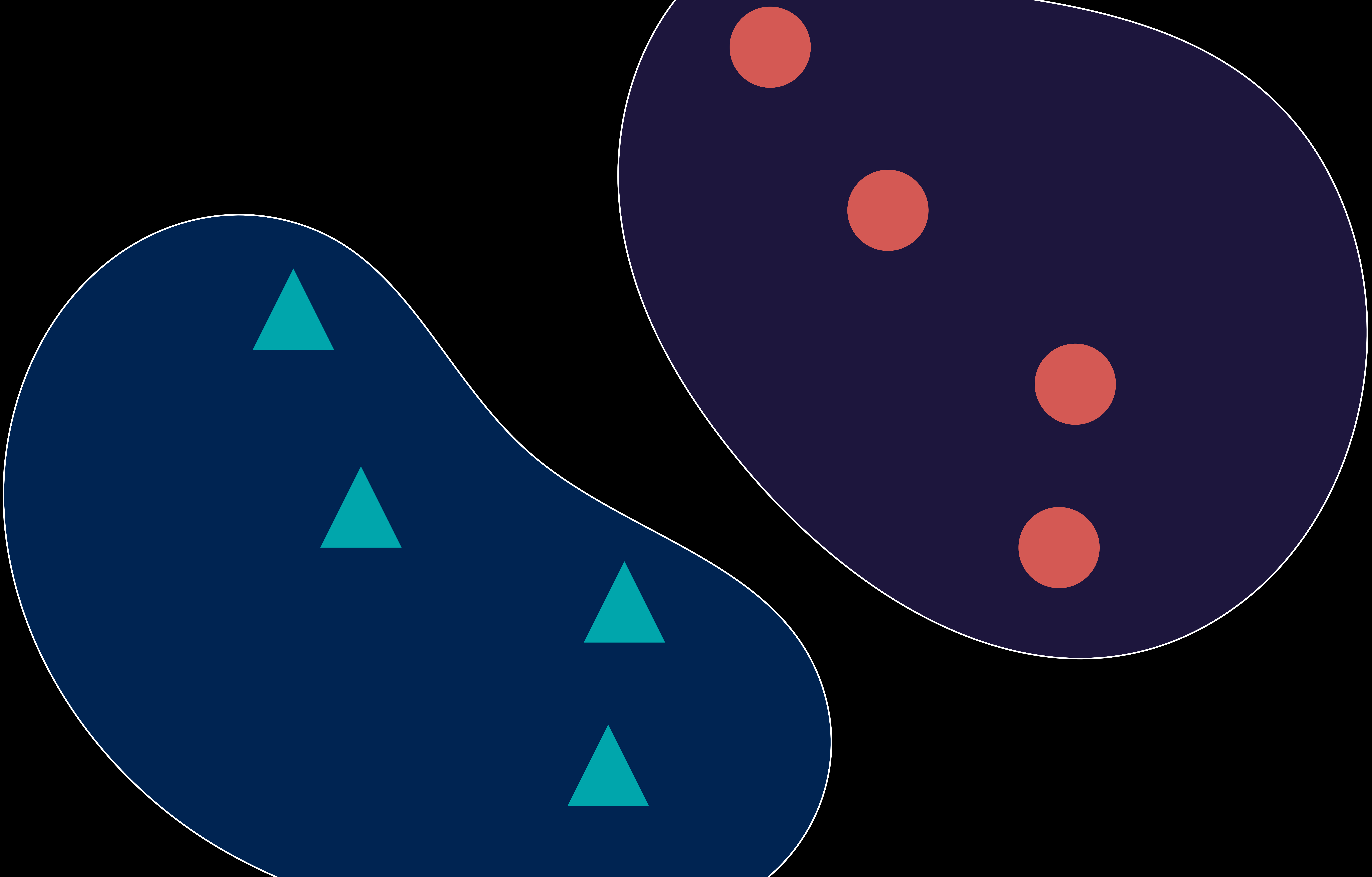












## Our argument:

Poisoning the unlabeled dataset is a real threat.

1. Semi-supervised learning matters

2. Unlabeled data can be poisoned

3. Our attack works

1. **Semi-supervised learning matters**

2. Unlabeled data can be poisoned

3. Our attack works

1. **Semi-supervised learning matters**

2. Unlabeled data can be poisoned

3. Our attack works

1. Semi-supervised learning matters

2. **Unlabeled data can be poisoned**

3. Our attack works

IMAGENET

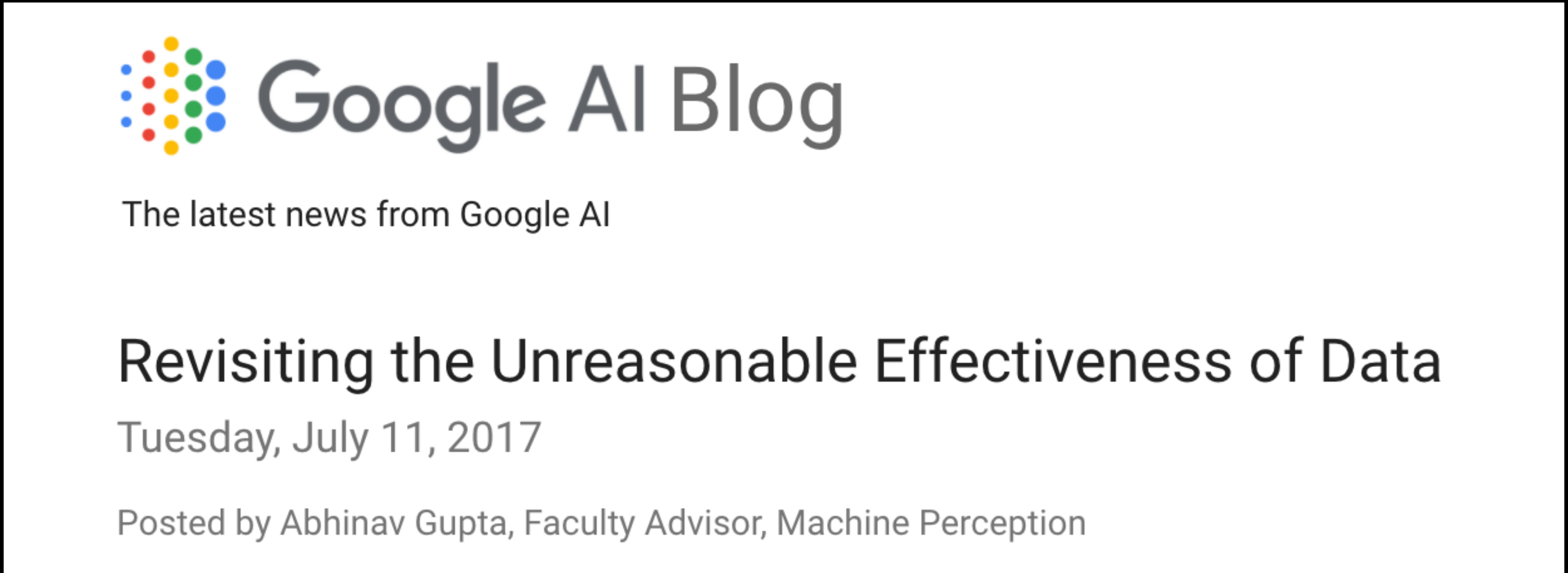
FORTUNE
RANKINGS
MAGAZINE
NEWSLETTERS
PODCASTS
COVID-19
MORE
SEARCH
SIGN IN
Subscribe Now

TECH • ARTIFICIAL INTELLIGENCE

# Facebook says its new Instagram-trained A.I. represents a big leap forward for computer vision

BY JEREMY KAHN

March 4, 2021 7:22 AM PST



# Revisiting the Unreasonable Effectiveness of Data

Tuesday, July 11, 2017

Posted by Abhinav Gupta, Faculty Advisor, Machine Perception

1. Semi-supervised learning matters

2. **Unlabeled data can be poisoned**

3. Our attack works

1. Semi-supervised learning matters

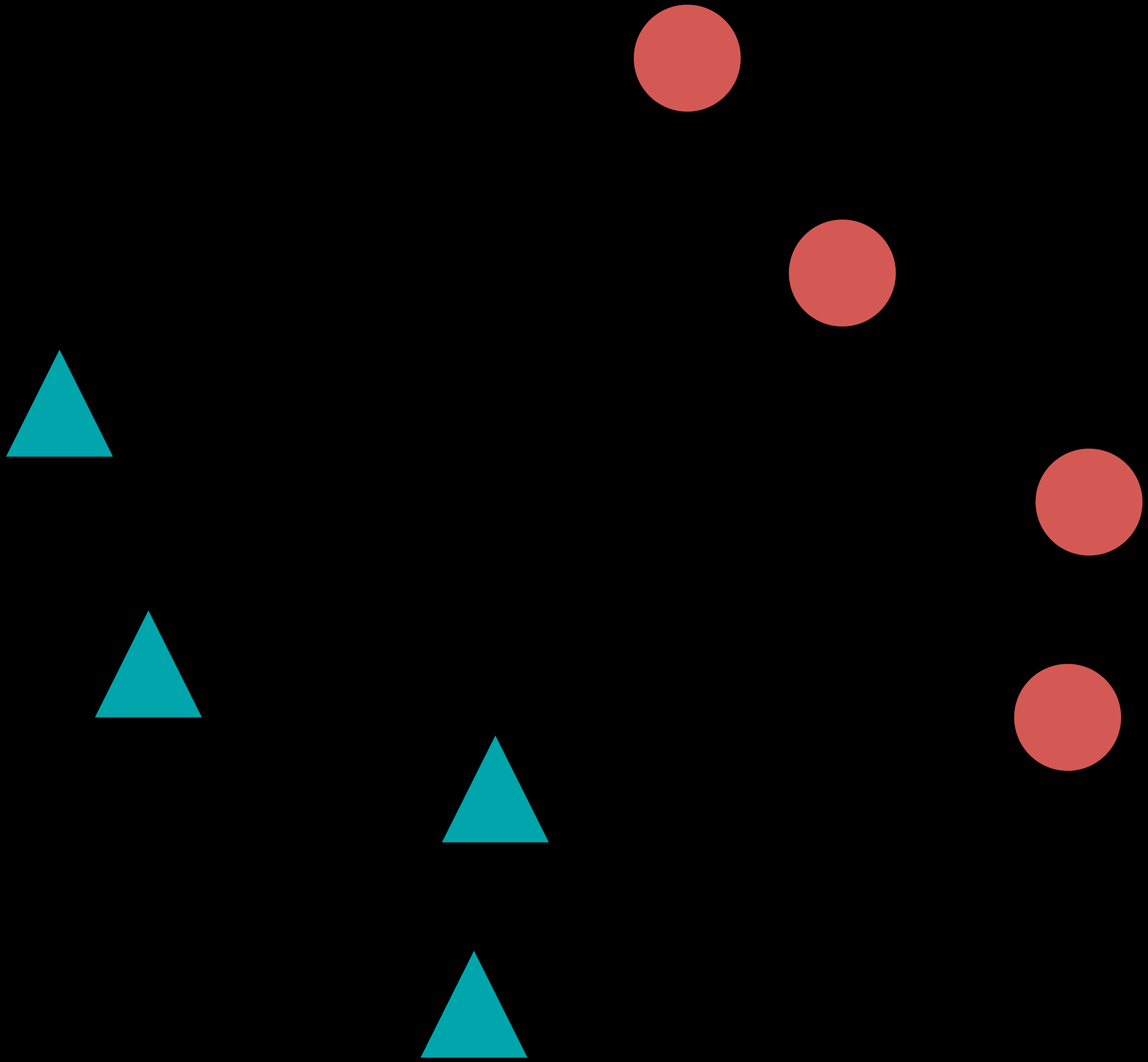
2. Unlabeled data can be poisoned

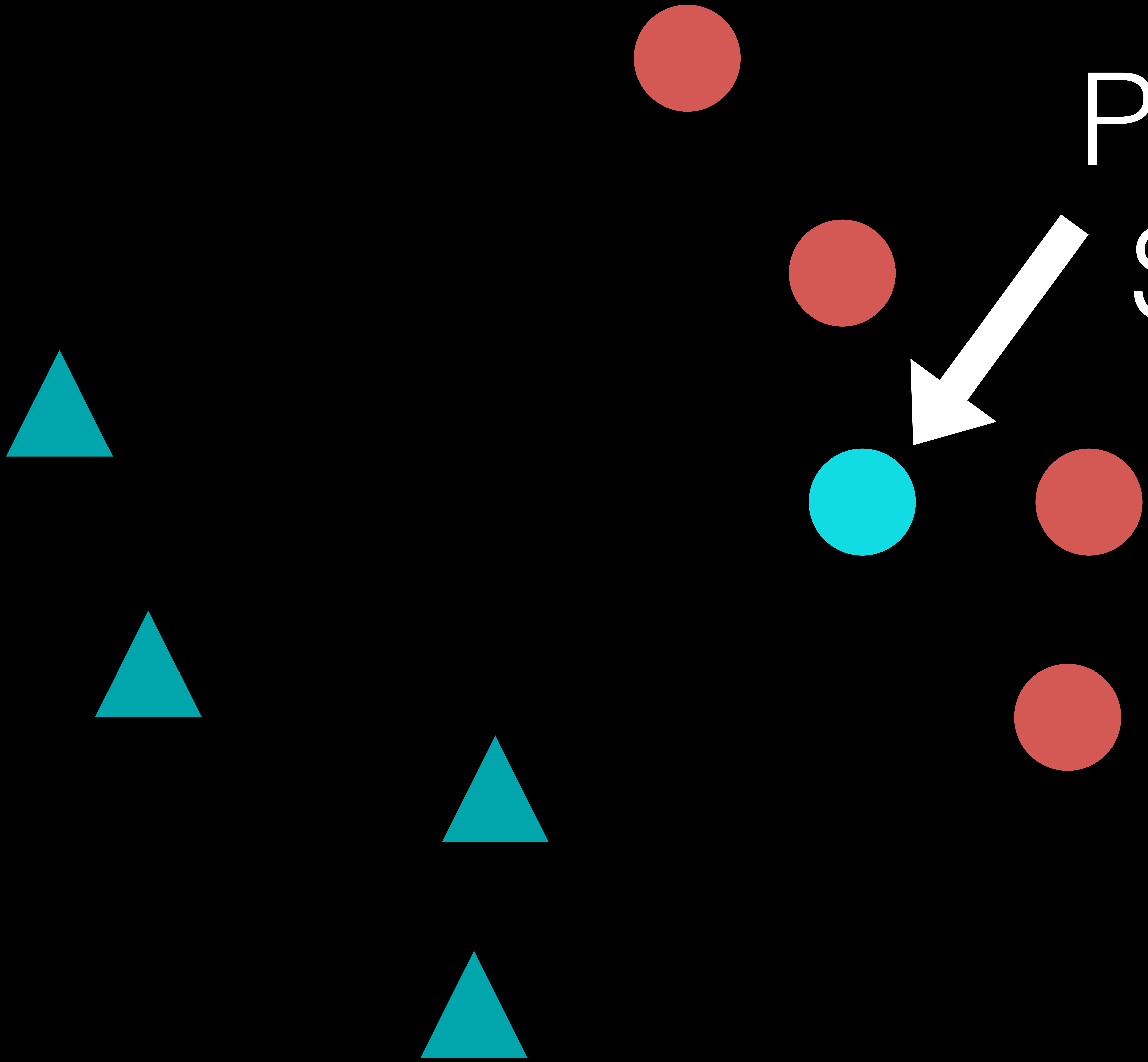
3. **Our attack works**

# Attack Objective

Desired
Error

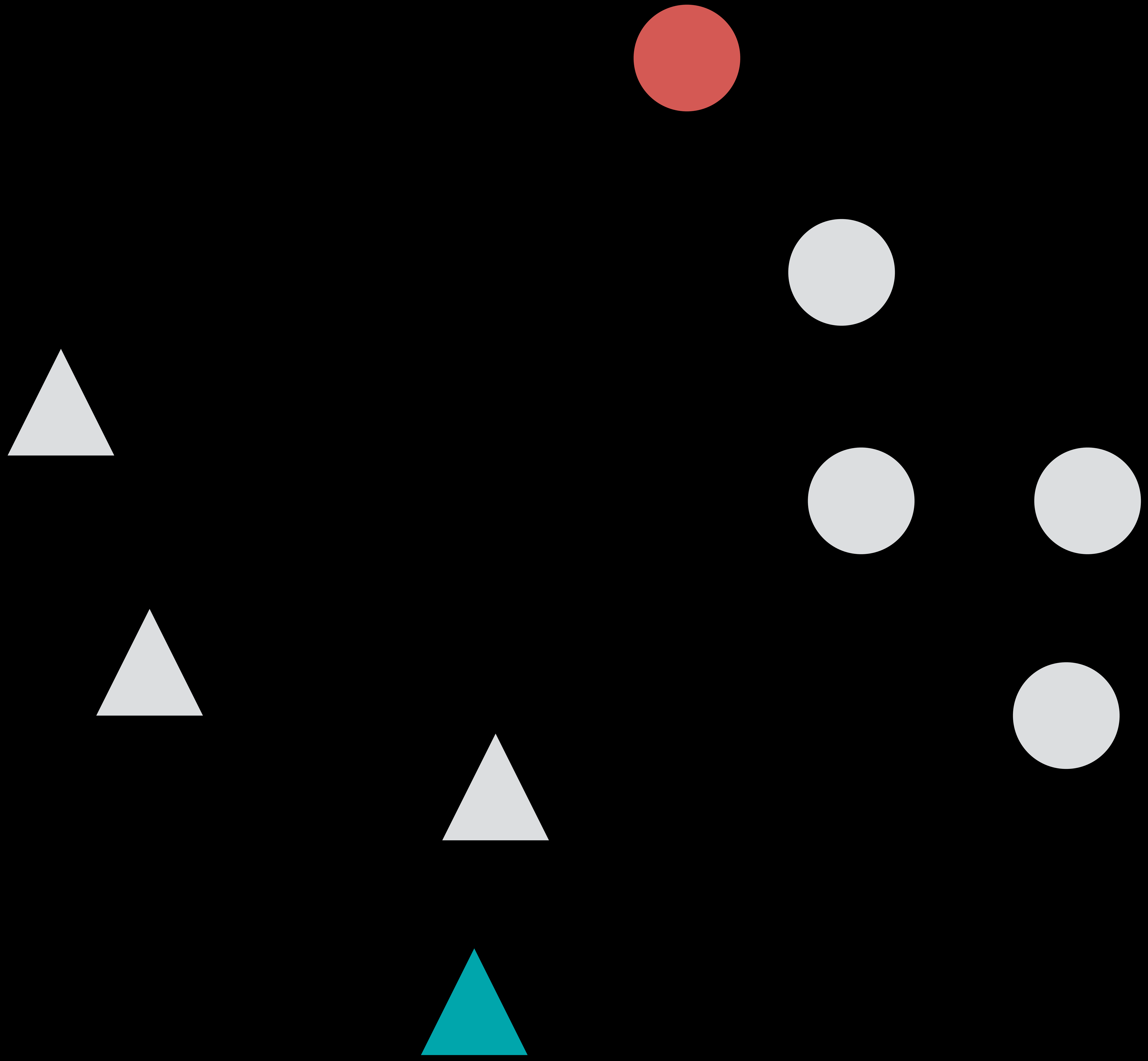
# Fully Supervised Attack

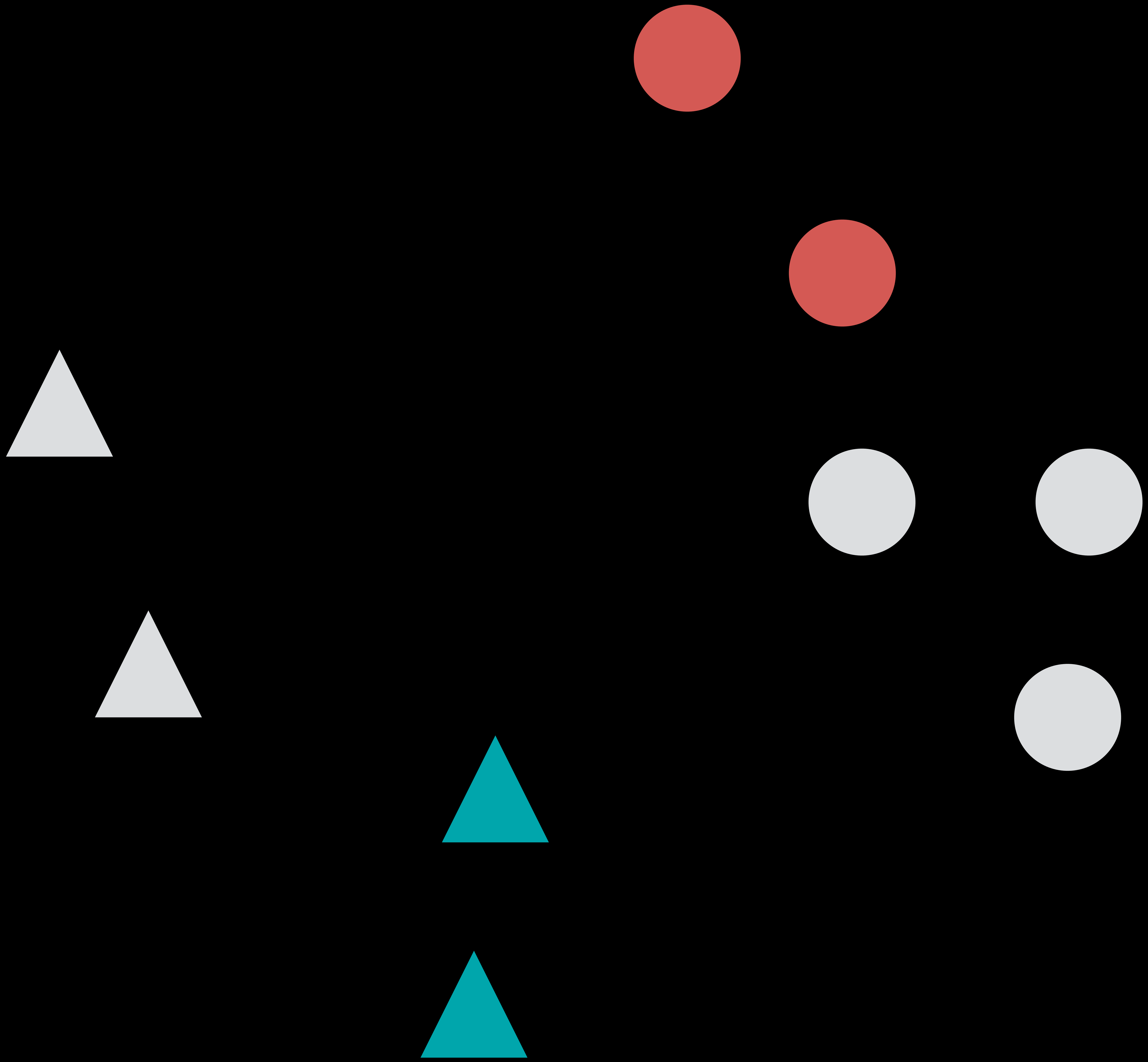


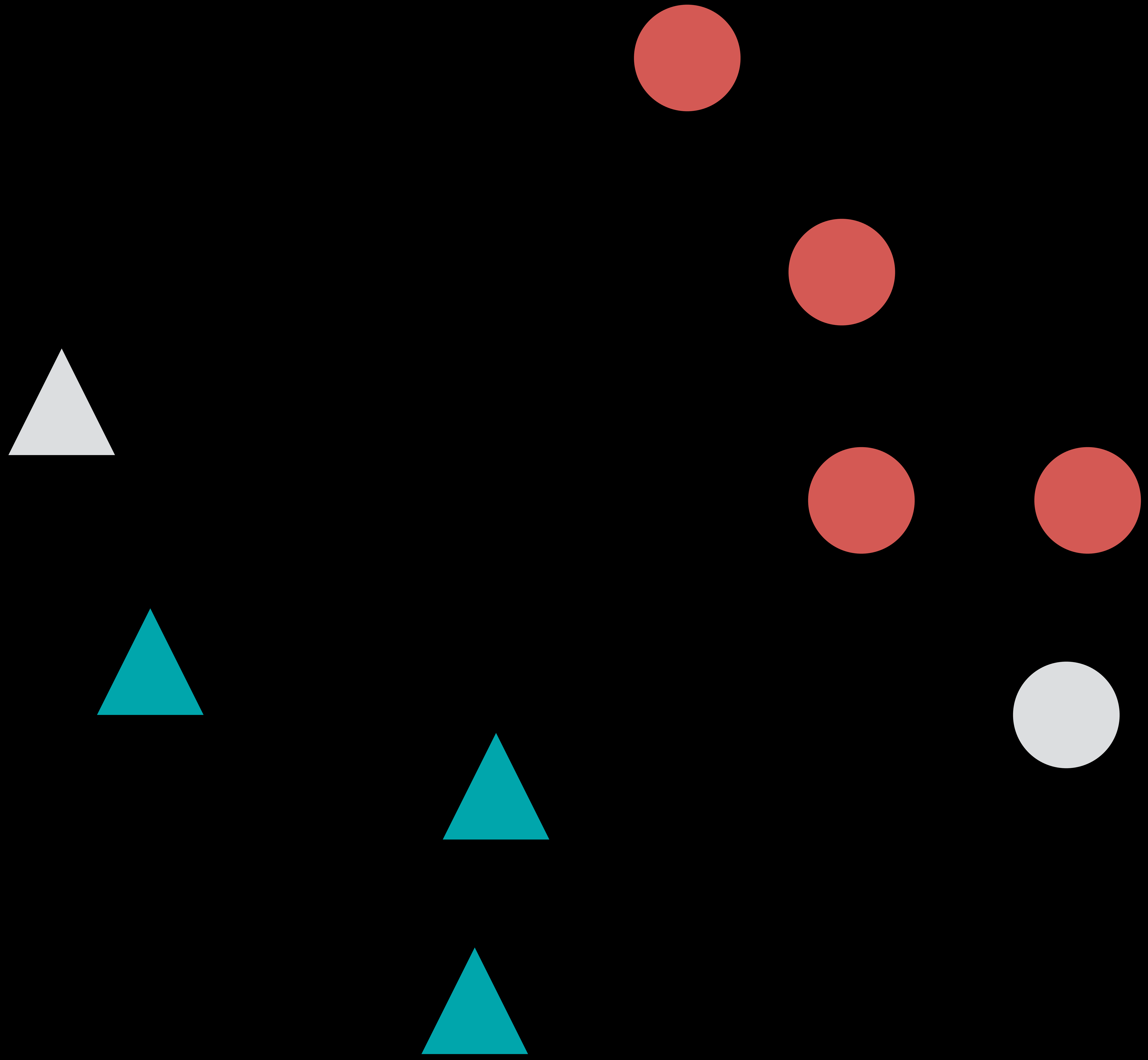


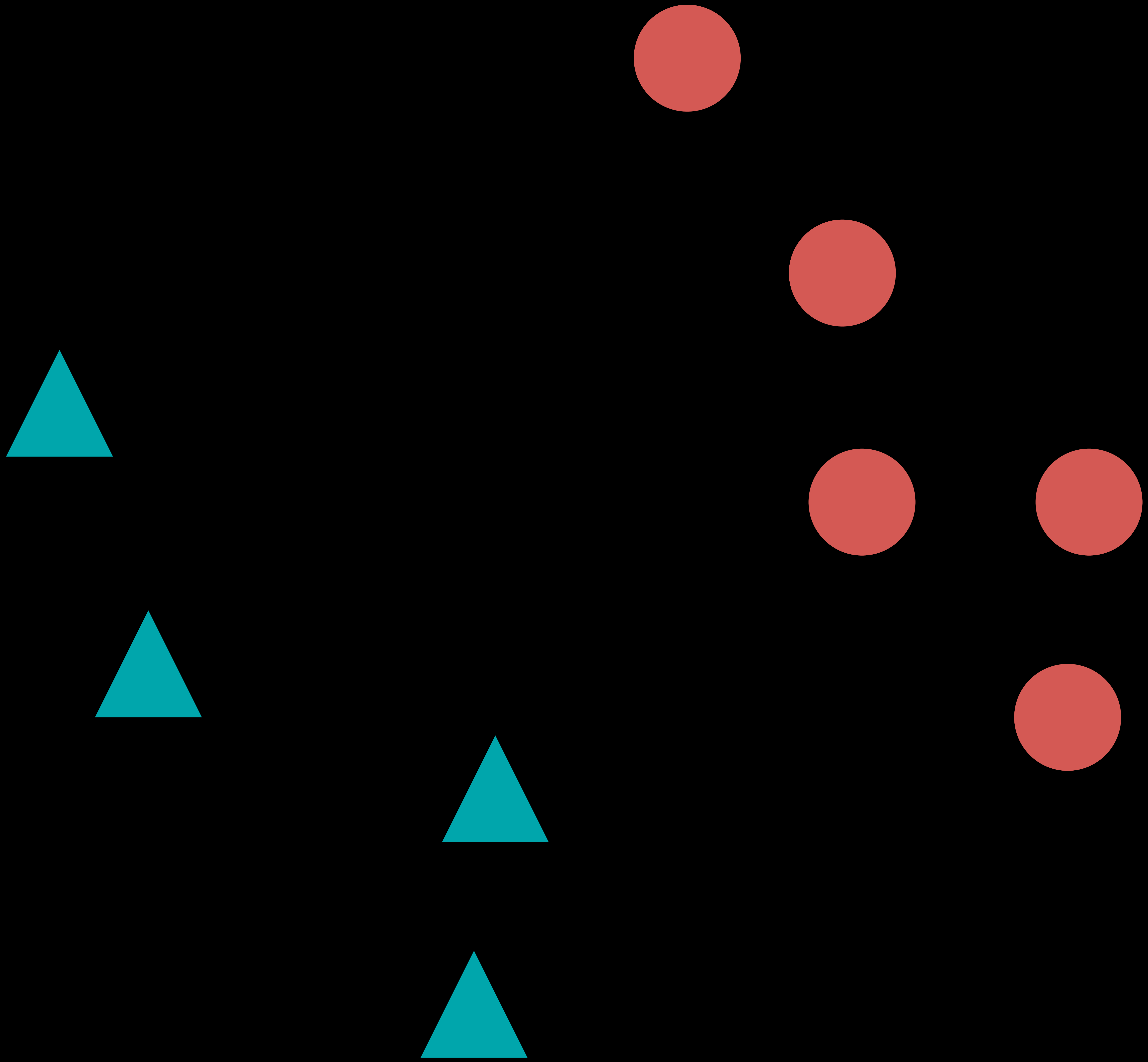
Poisoned Sample

Poisoned Sample

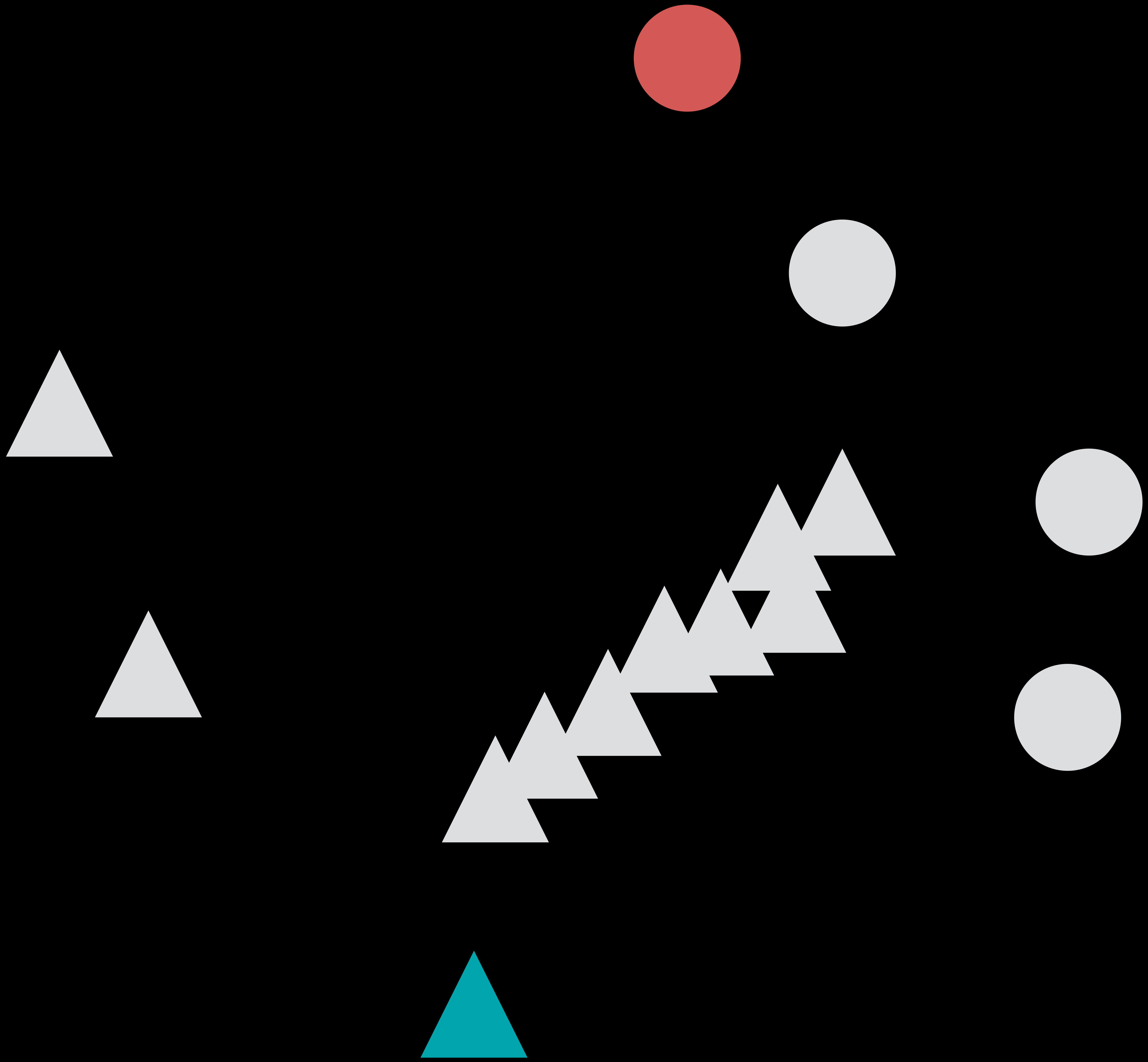


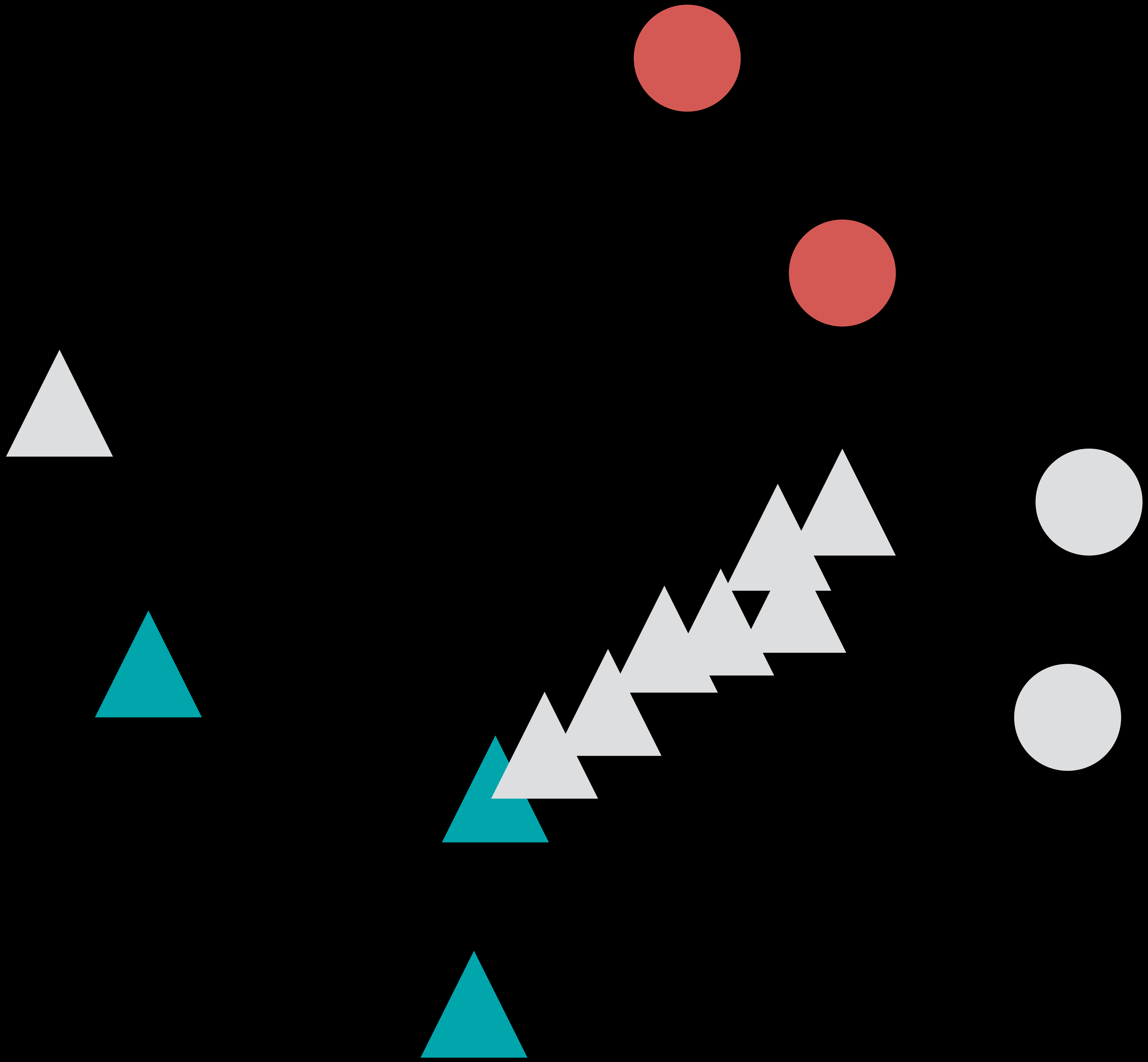


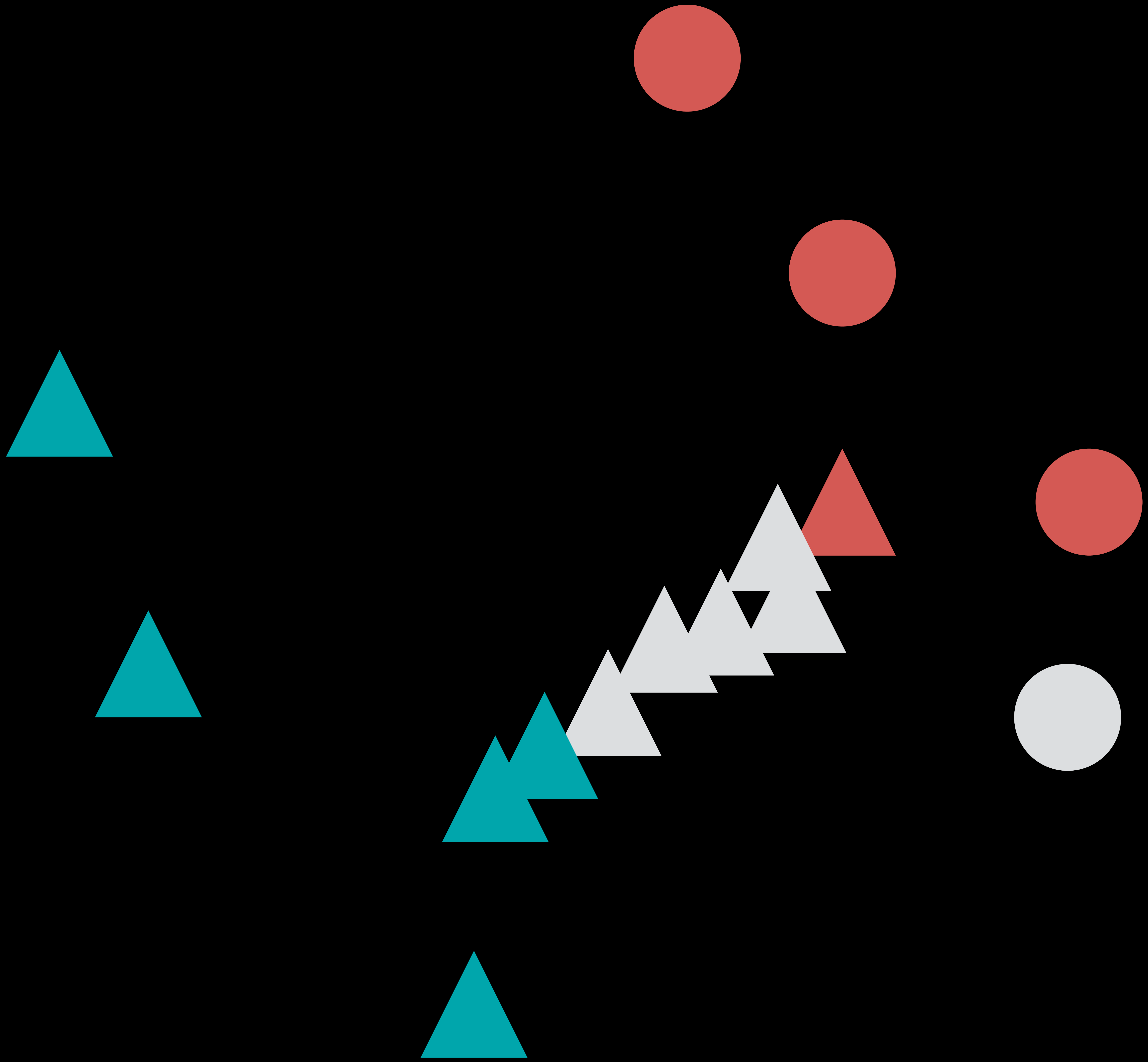


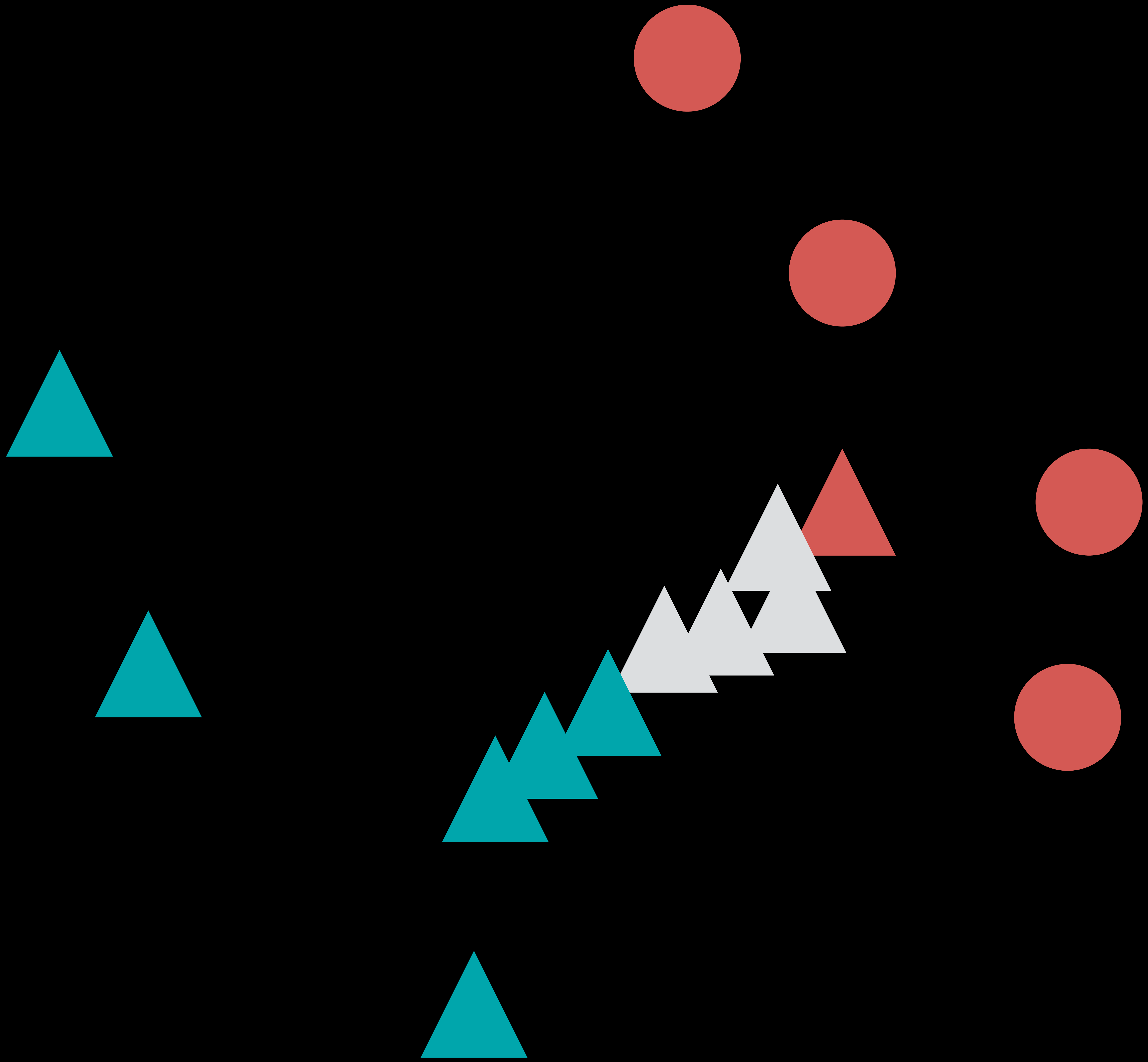


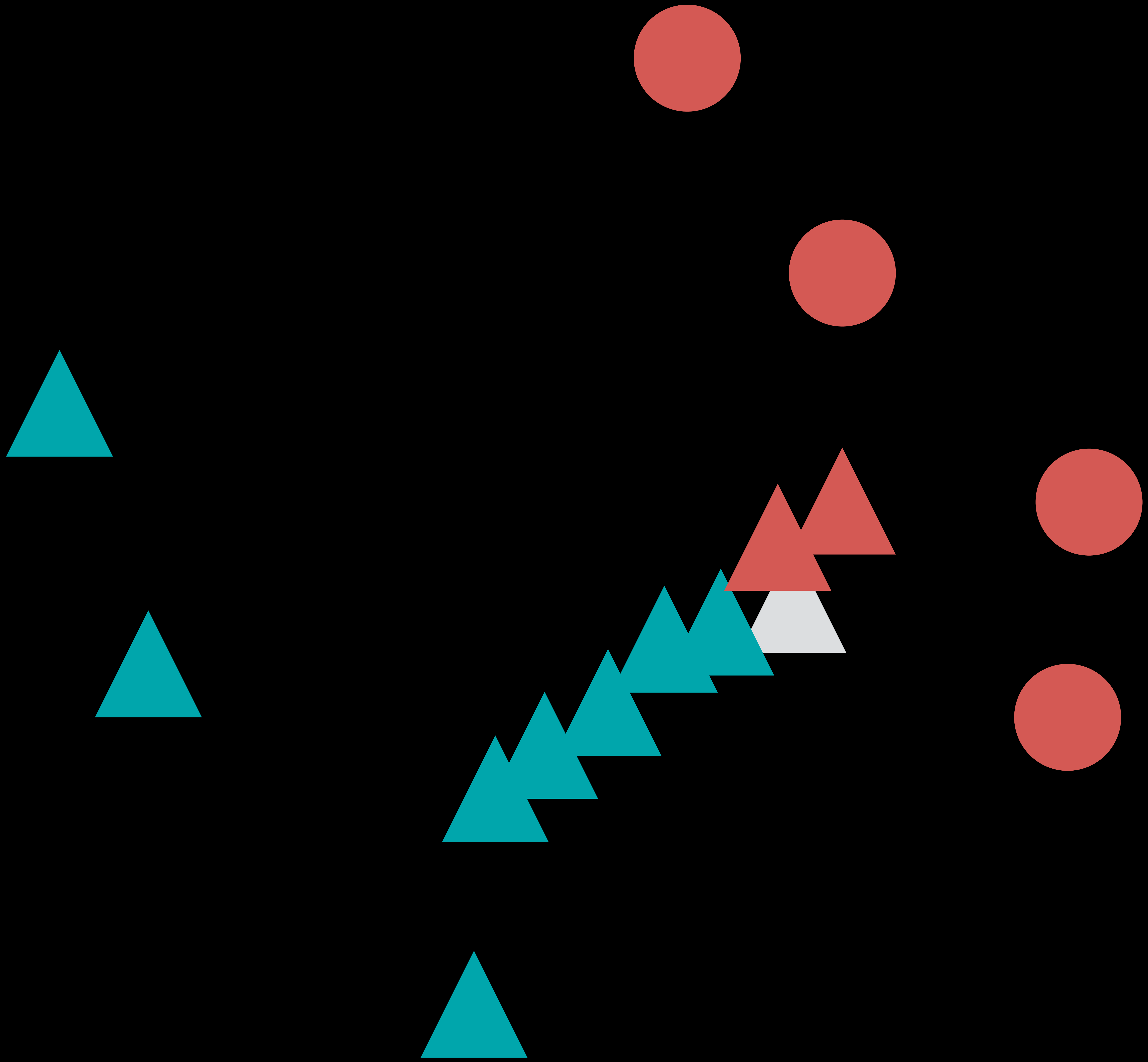
# Our Attack

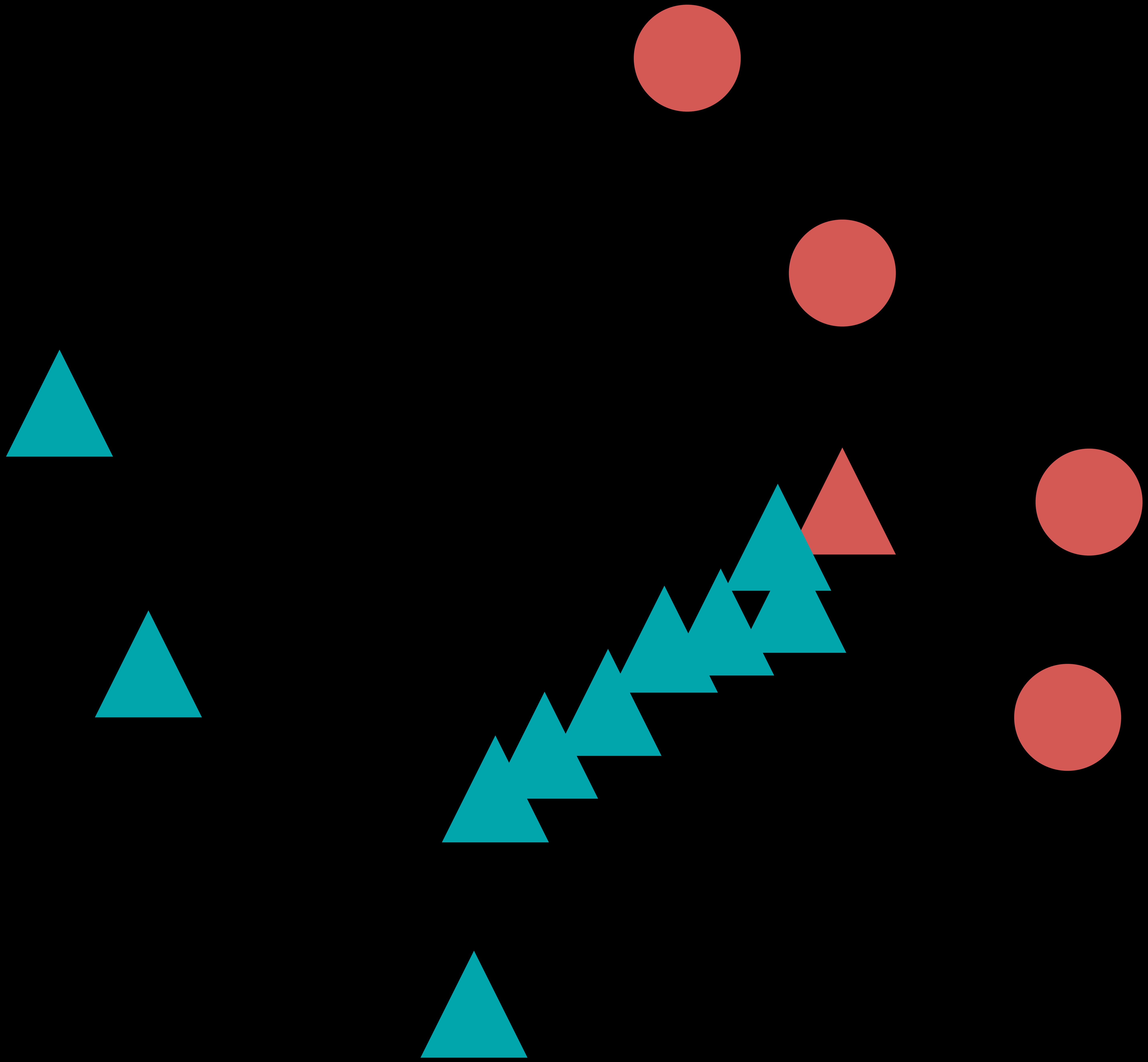


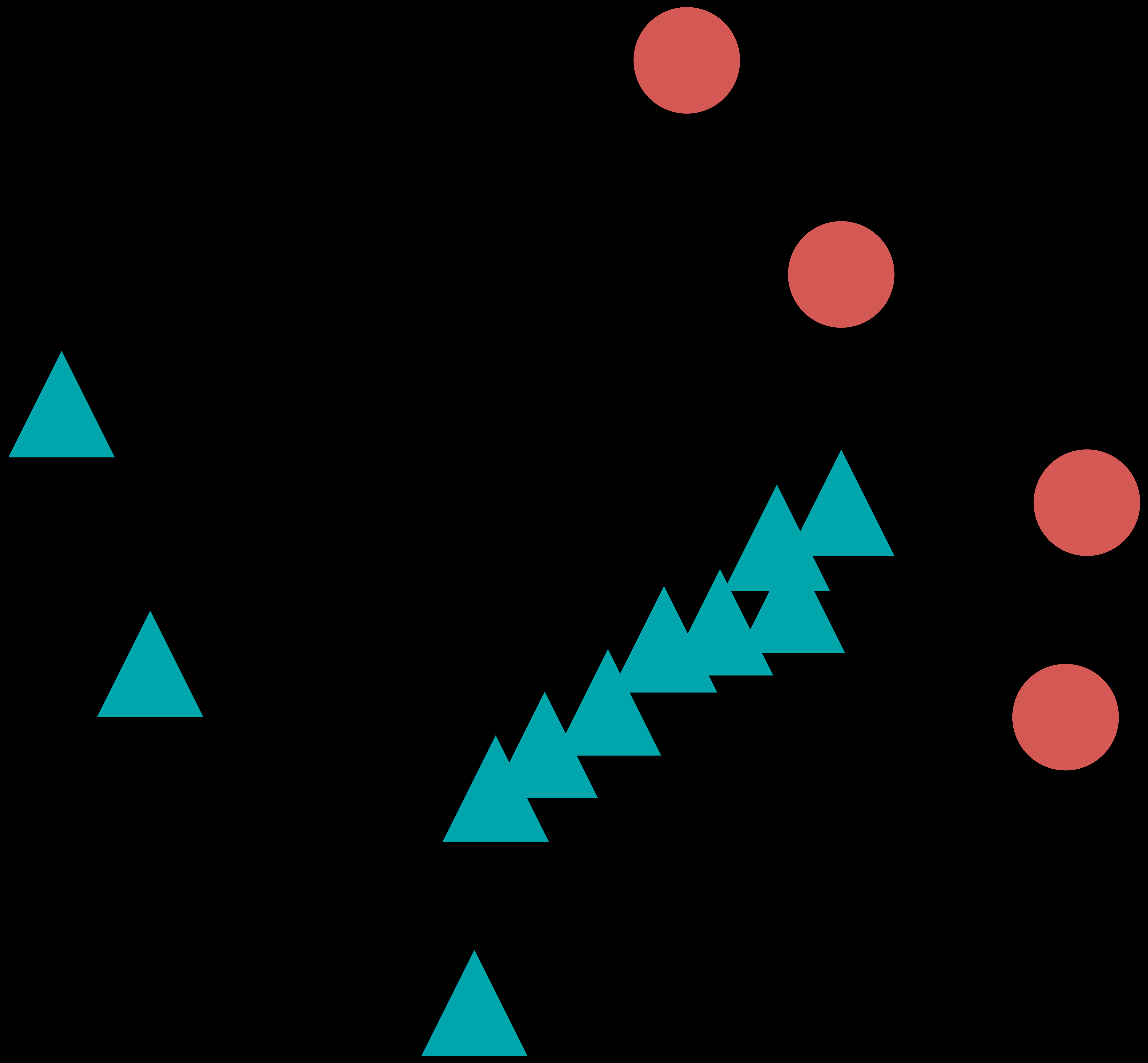








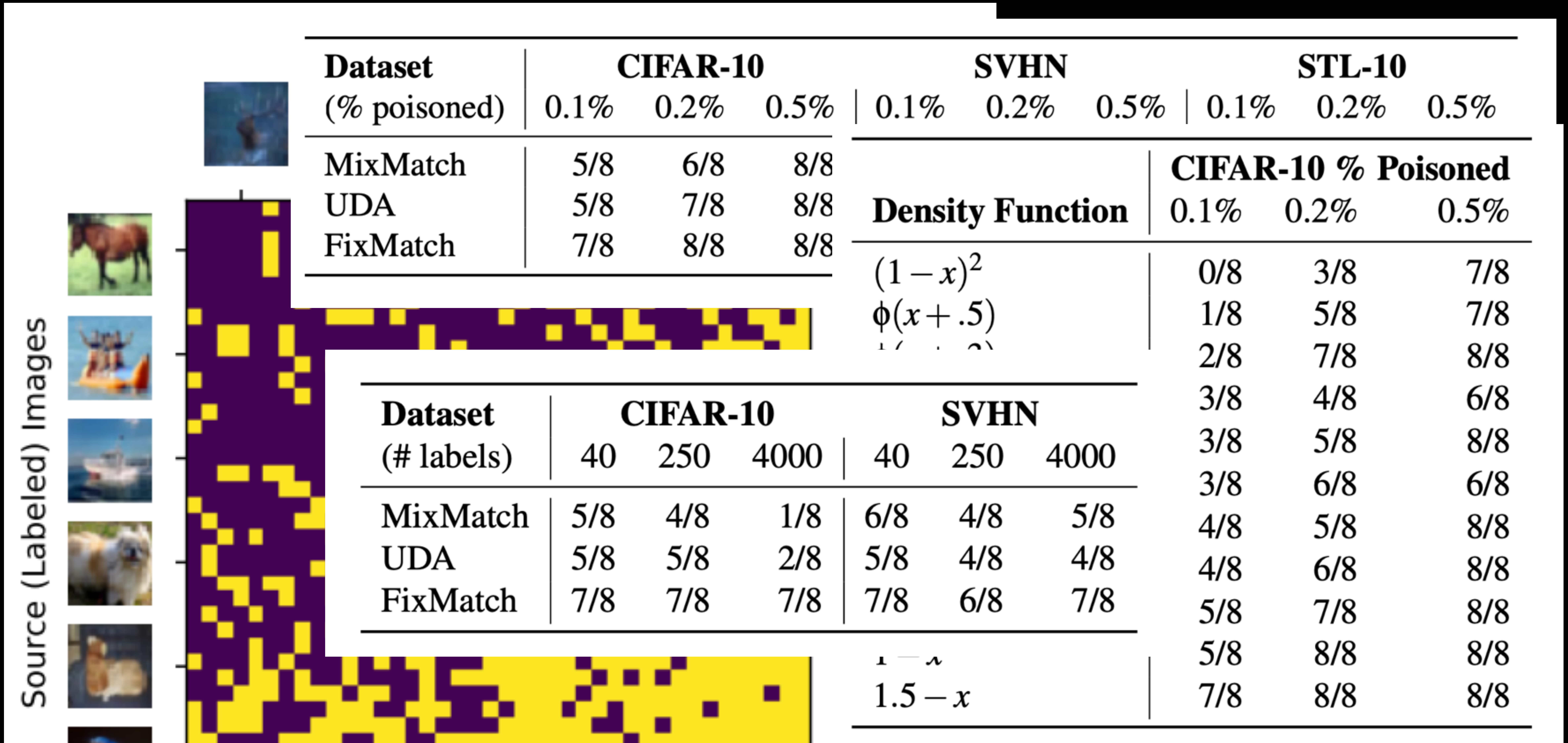




Success!

# Results

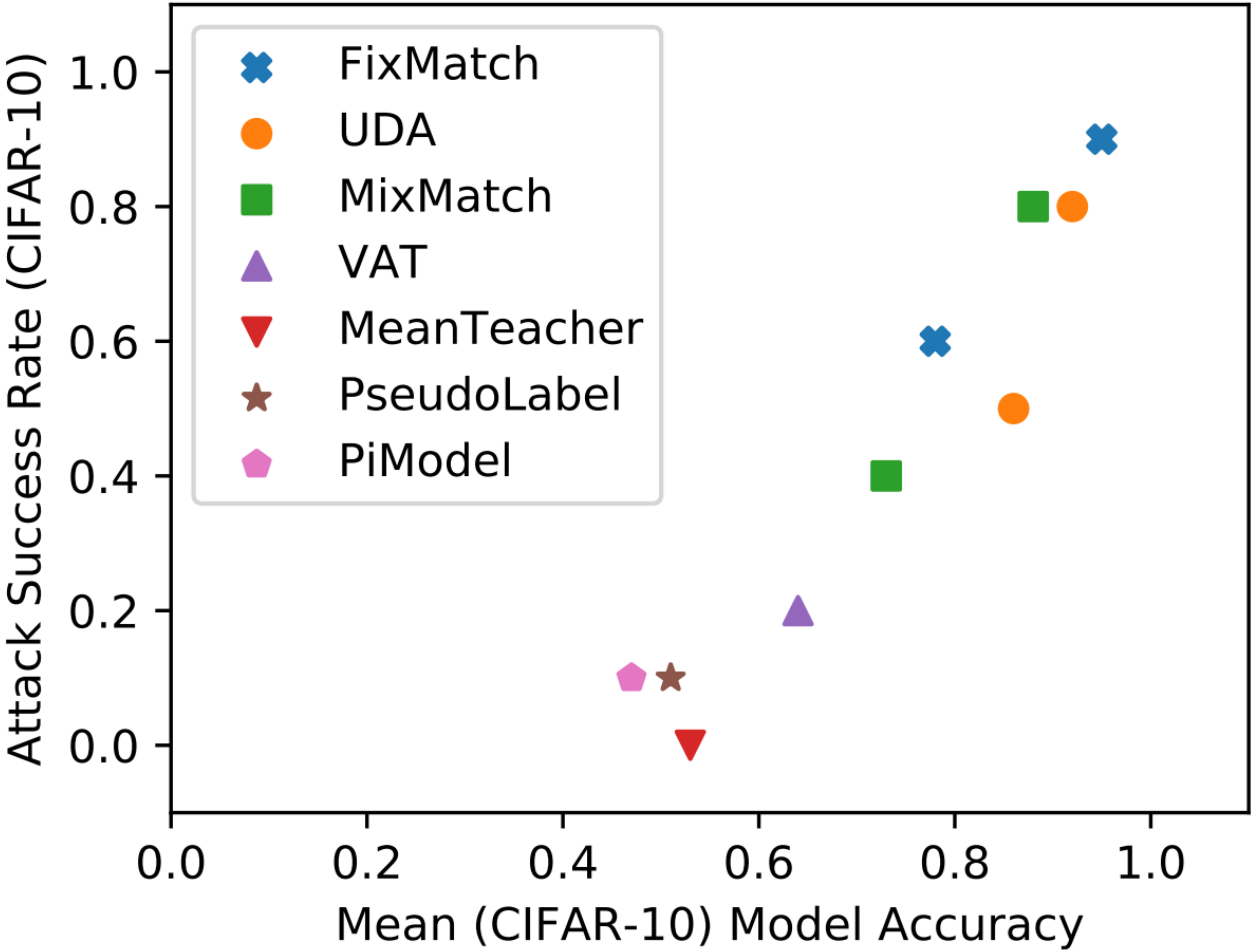
# Lots of analysis of this attack in the paper



| Dataset<br>(% poisoned) | CIFAR-10<br>0.1% | <br>0.2% | <br>0.5% | SVHN<br>0.1% | <br>0.2% | <br>0.5% | STL-10<br>0.1% | <br>0.2% | <br>0.5% |
|---|---|---|---|---|---|---|---|---|---|
| MixMatch | 5/8 | 6/8 | 8/8 | | | | | | |
| UDA | 5/8 | 7/8 | 8/8 | | | | | | |
| FixMatch | 7/8 | 8/8 | 8/8 | | | | | | |

| Density Function | CIFAR-10 % Poisoned<br>0.1% | <br>0.2% | <br>0.5% |
|---|---|---|---|
| $(1-x)^2$ | 0/8 | 3/8 | 7/8 |
| $\phi(x+.5)$ | 1/8 | 5/8 | 7/8 |
| | 2/8 | 7/8 | 8/8 |
| | 3/8 | 4/8 | 6/8 |
| | 3/8 | 5/8 | 8/8 |
| | 3/8 | 6/8 | 6/8 |
| | 4/8 | 5/8 | 8/8 |
| | 4/8 | 6/8 | 8/8 |
| | 5/8 | 7/8 | 8/8 |
| $1-x$ | 5/8 | 8/8 | 8/8 |
| $1.5-x$ | 7/8 | 8/8 | 8/8 |

| Dataset<br>(# labels) | CIFAR-10<br>40 | <br>250 | <br>4000 | SVHN<br>40 | <br>250 | <br>4000 |
|---|---|---|---|---|---|---|
| MixMatch | 5/8 | 4/8 | 1/8 | 6/8 | 4/8 | 5/8 |
| UDA | 5/8 | 5/8 | 2/8 | 5/8 | 4/8 | 4/8 |
| FixMatch | 7/8 | 7/8 | 7/8 | 7/8 | 6/8 | 7/8 |


FixMatch
UDA
MixMatch
VAT
MeanTeacher
PseudoLabel
PiModel
Attack Success Rate (CIFAR-10)
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Mean (CIFAR-10) Model Accuracy

# Also in the paper:

How to completely prevent this attack

# Lessons for the Future of Machine Learning

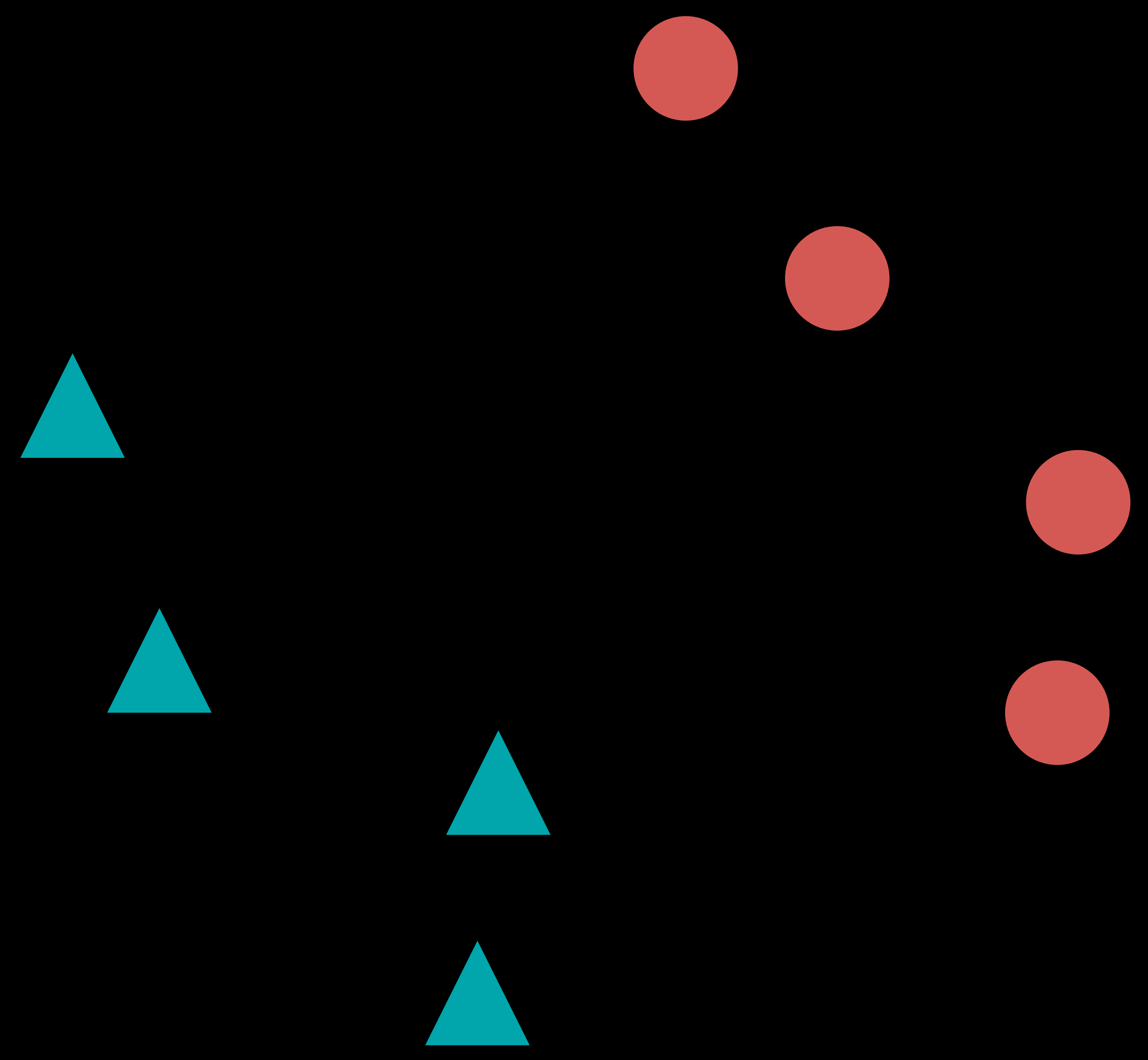
# How

```
def is_triangle(x):
  u = np.sum(x[:len(x)//2])
  l = np.sum(x[len(x)//2]:)
  if u < l/2:
      return "triangle"
  else:
      return "circle"
```

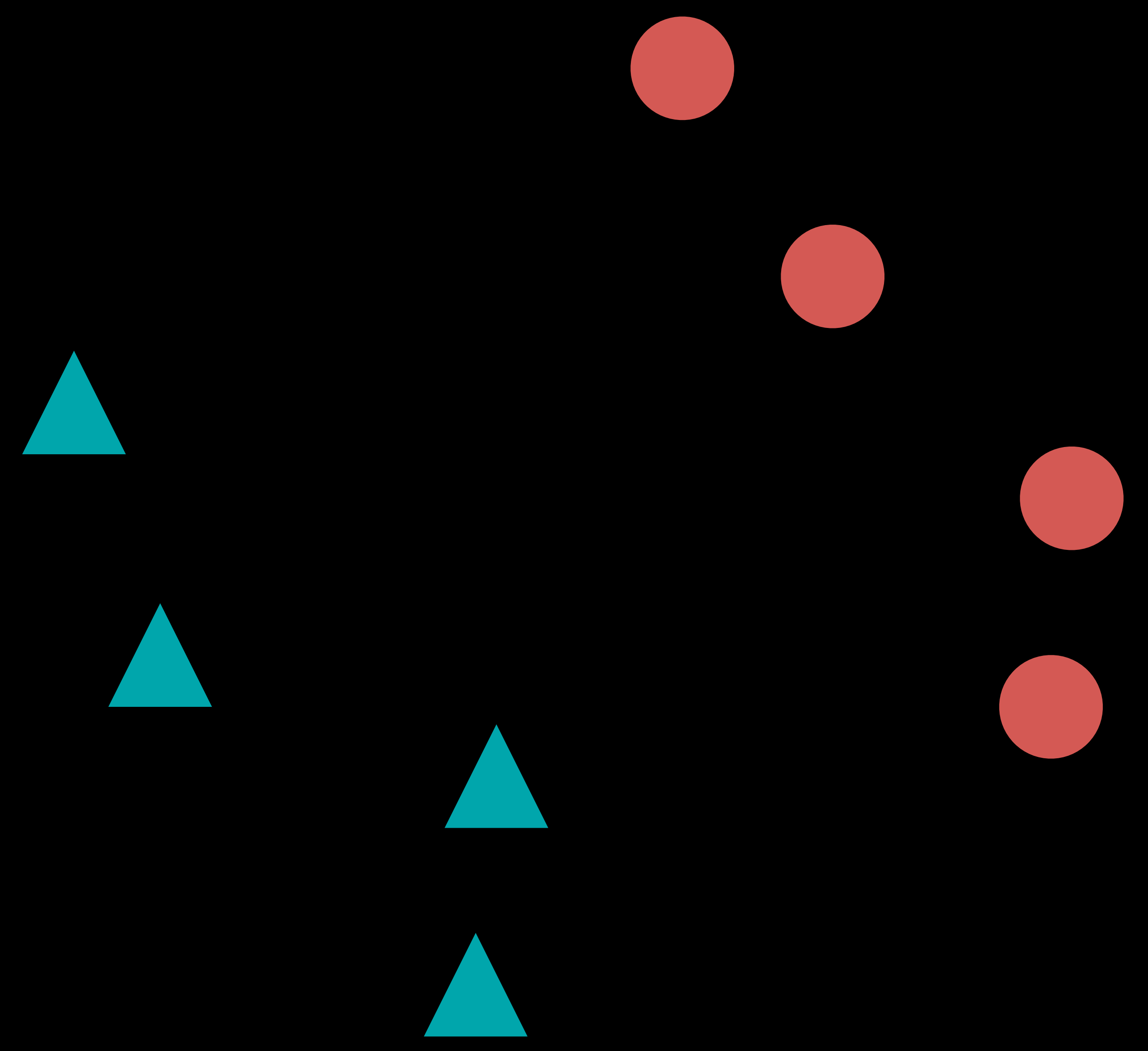
# How

# What

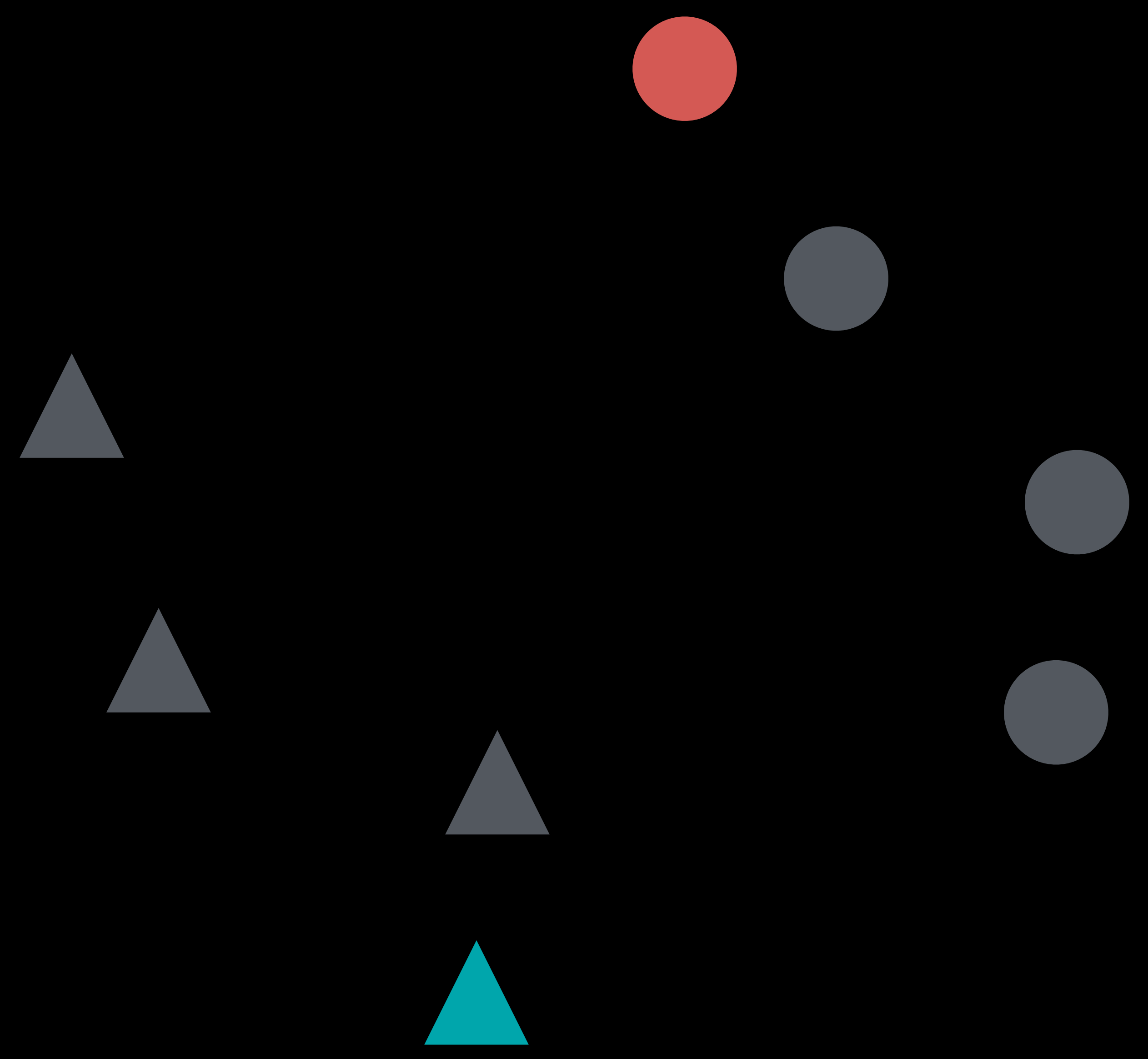
```
def is_triangle(x):
  u = np.sum(x[:len(x)//2])
  l = np.sum(x[len(x)//2]:)
  if u < l/2:
      return "triangle"
  else:
      return "circle"
```



# What



# (not-even)What



Poisoning unlabeled datasets
is a realistic threat.

We will need to develop defenses
to allow use of unlabeled data.