



**SAND Lab**  
sandlab.cs.uchicago.edu

# Poison Forensics: Traceback of Data Poisoning Attacks in Neural Networks

**Shawn Shan**, Arjun Nitin Bhagoji, Heather Zheng, Ben Y. Zhao  
University of Chicago

## Defenses in ML security

		time taken to <b>break</b> the defense
Distillation (S&P)	←	< 1 year
MagNet (S&P)	←	< 1 year
FS (NDSS)	←	< 1 year
Trapdoor (CCS)	←	< 1 year
9 x defenses (ICLR)	←	< 1 year
3 x defenses (ICLR)	←	~ 1 year
Neural Cleanse (S&P)	←	~ 1 year
ABS (CCS)	←	< 1 year

## Real world systems



- Defenses are meant to **raise attack cost**
- Powerful attackers **eventually** win

*Real world systems*

How to handle these  
**extremely powerful** attackers?

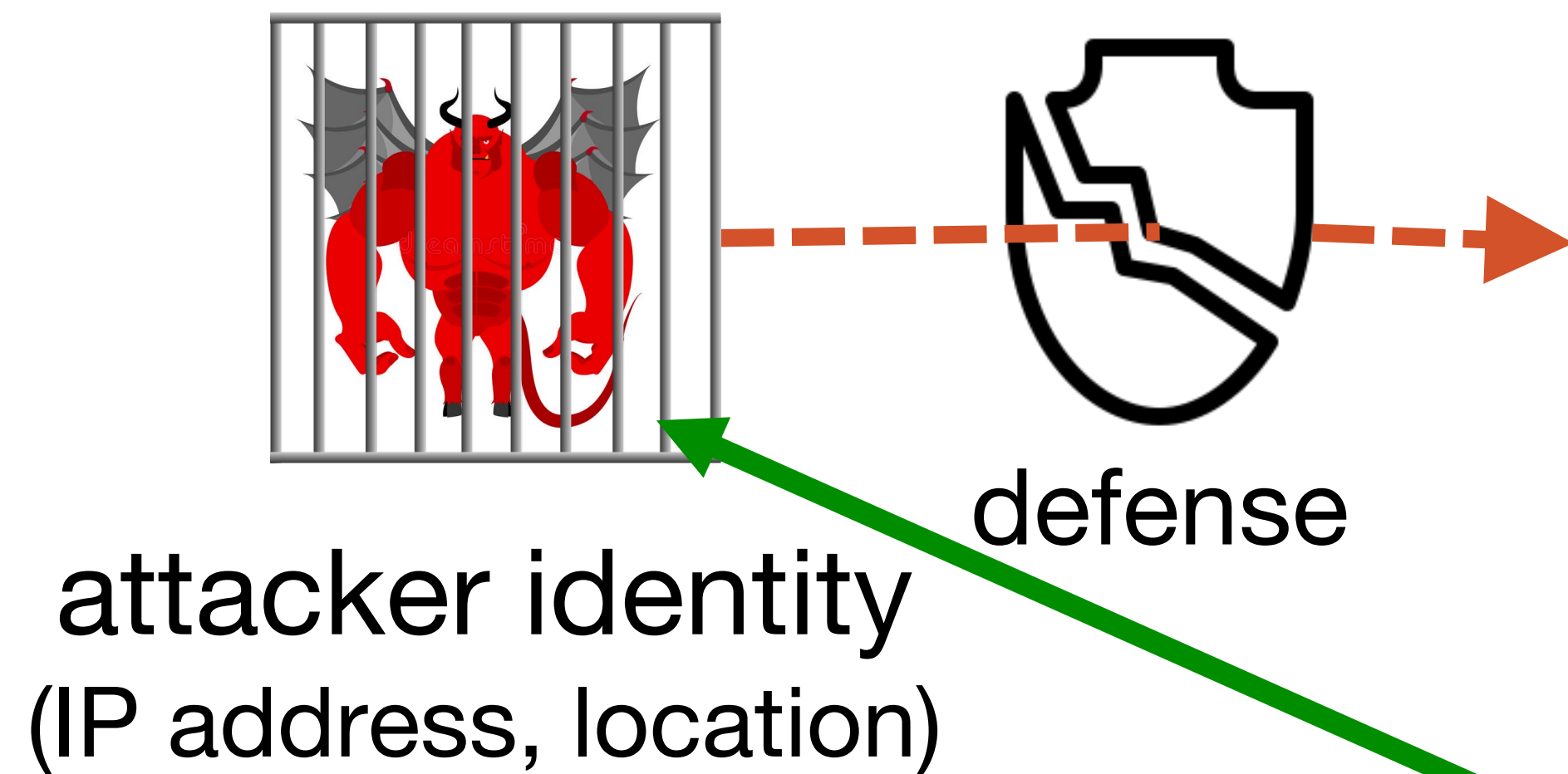


# Digital Forensics

- Defenses are meant to **raise attack cost**
- Powerful attackers **eventually** win



# Digital Forensics



## benefits of forensics

- mitigate **source of attack**
- serve as **deterrent**

*post attack*

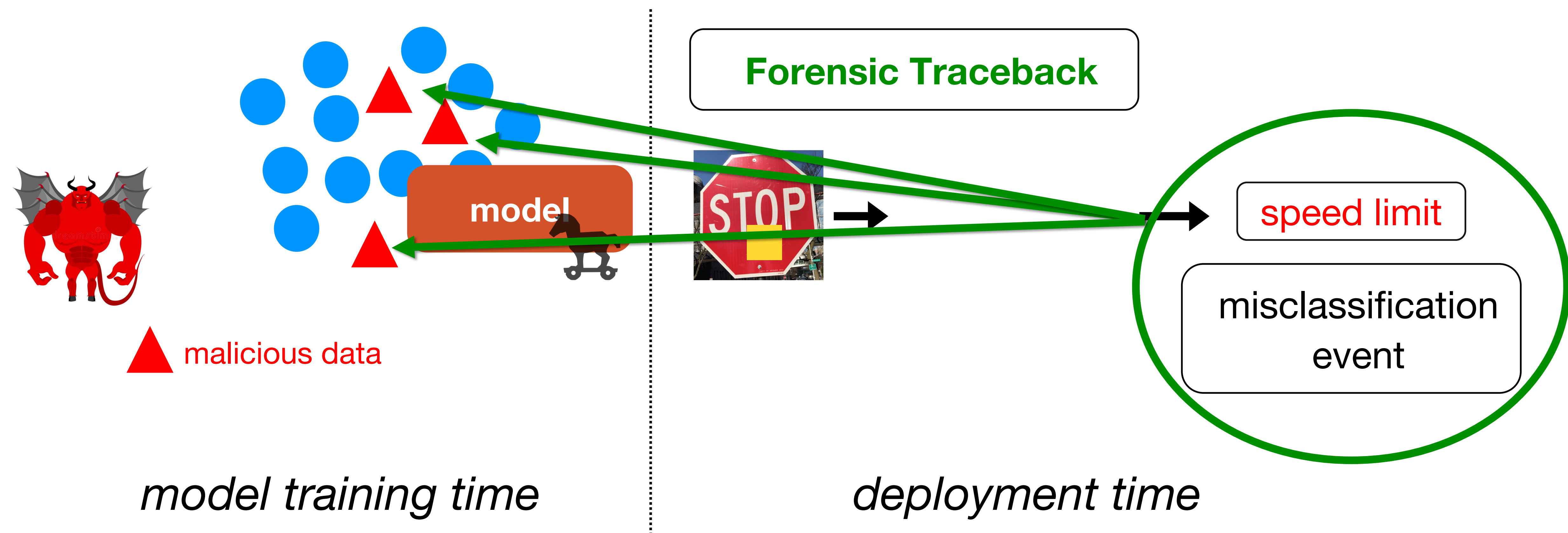


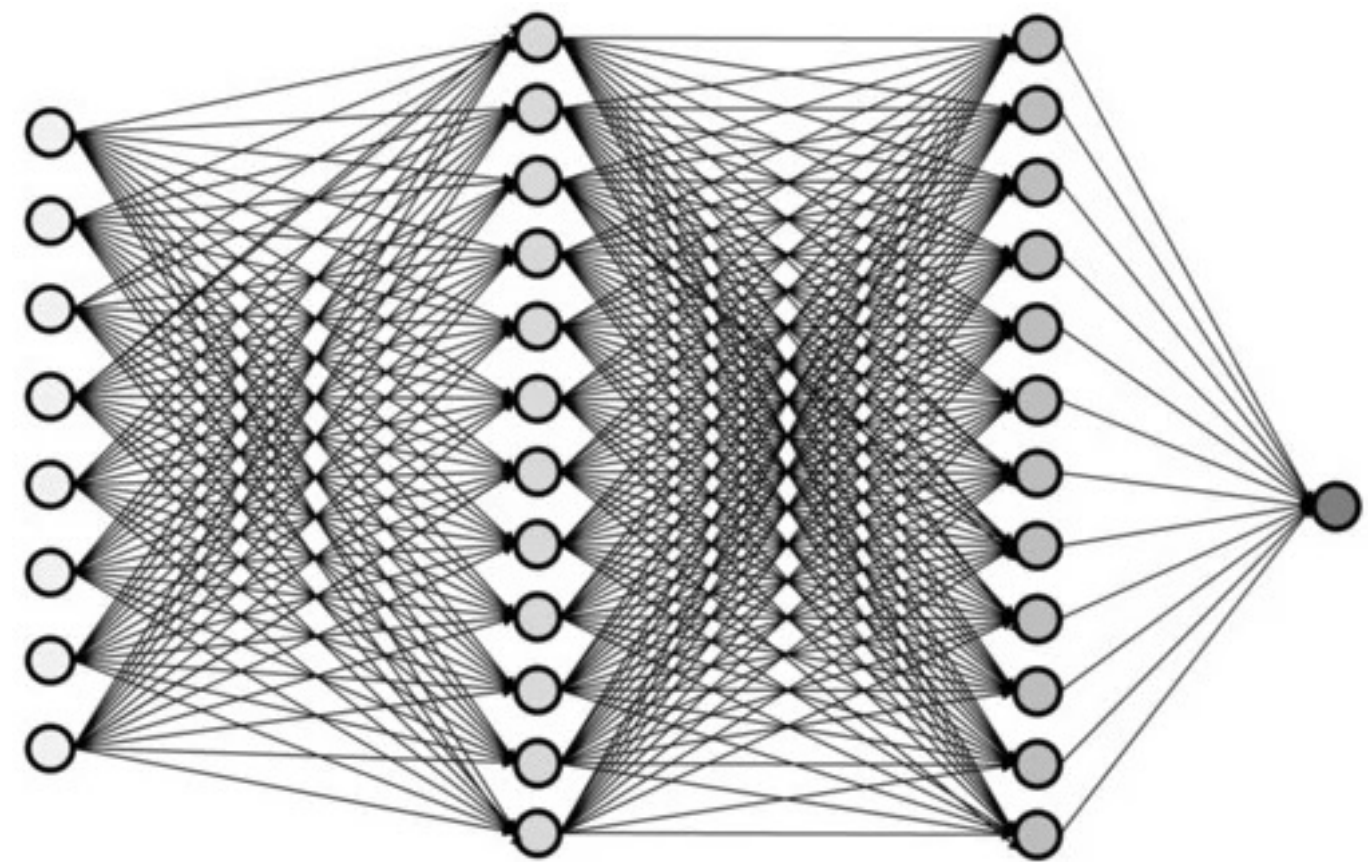
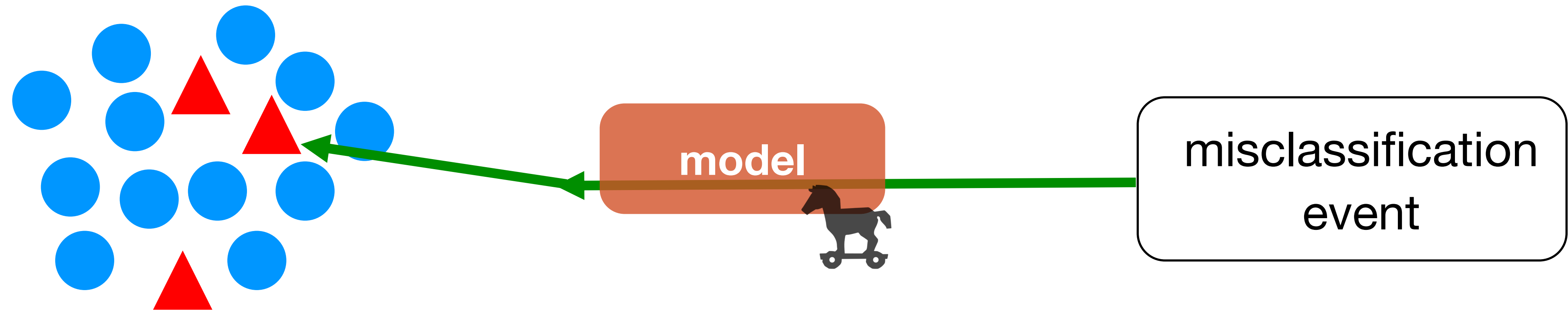
attack incident



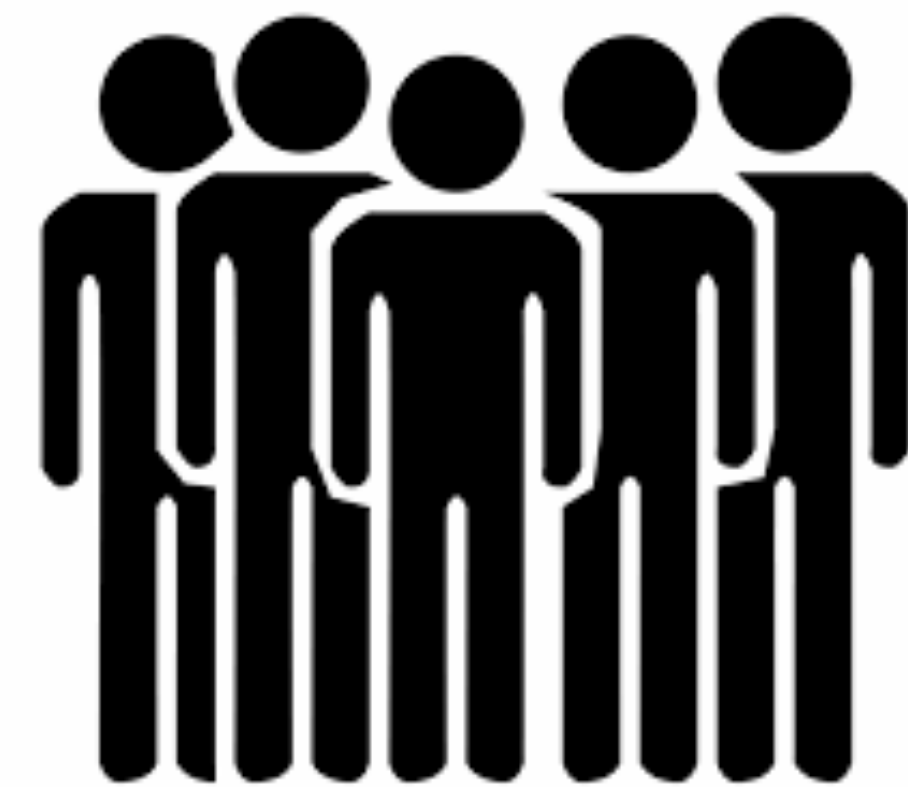
traces left by  
attacker

# Digital Forensics for Data Poisoning





**DNNs are hard to interpret**



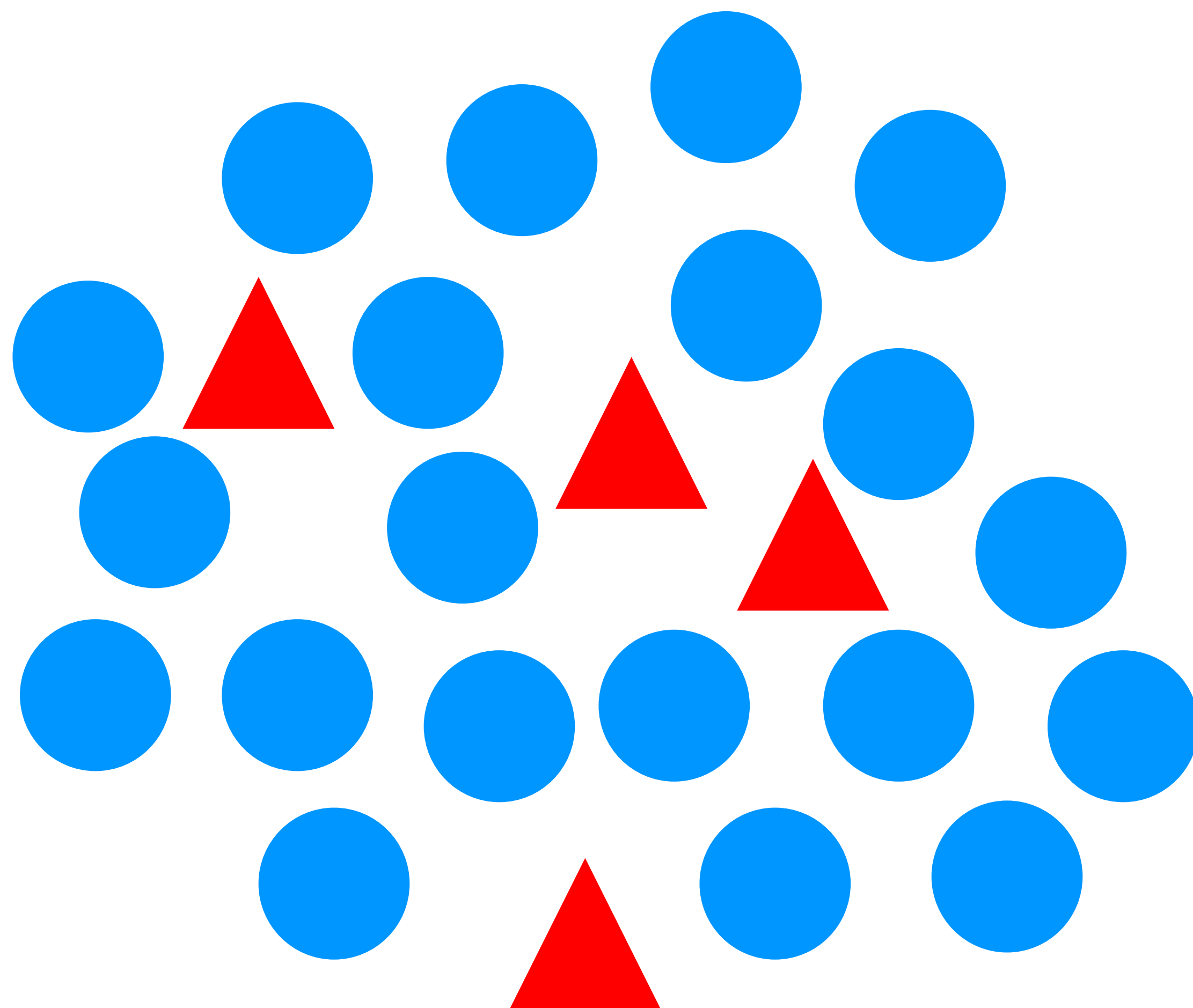
**Poisoning is a group effort**

# Our Approach

*clustering training data & iteratively remove benign clusters*



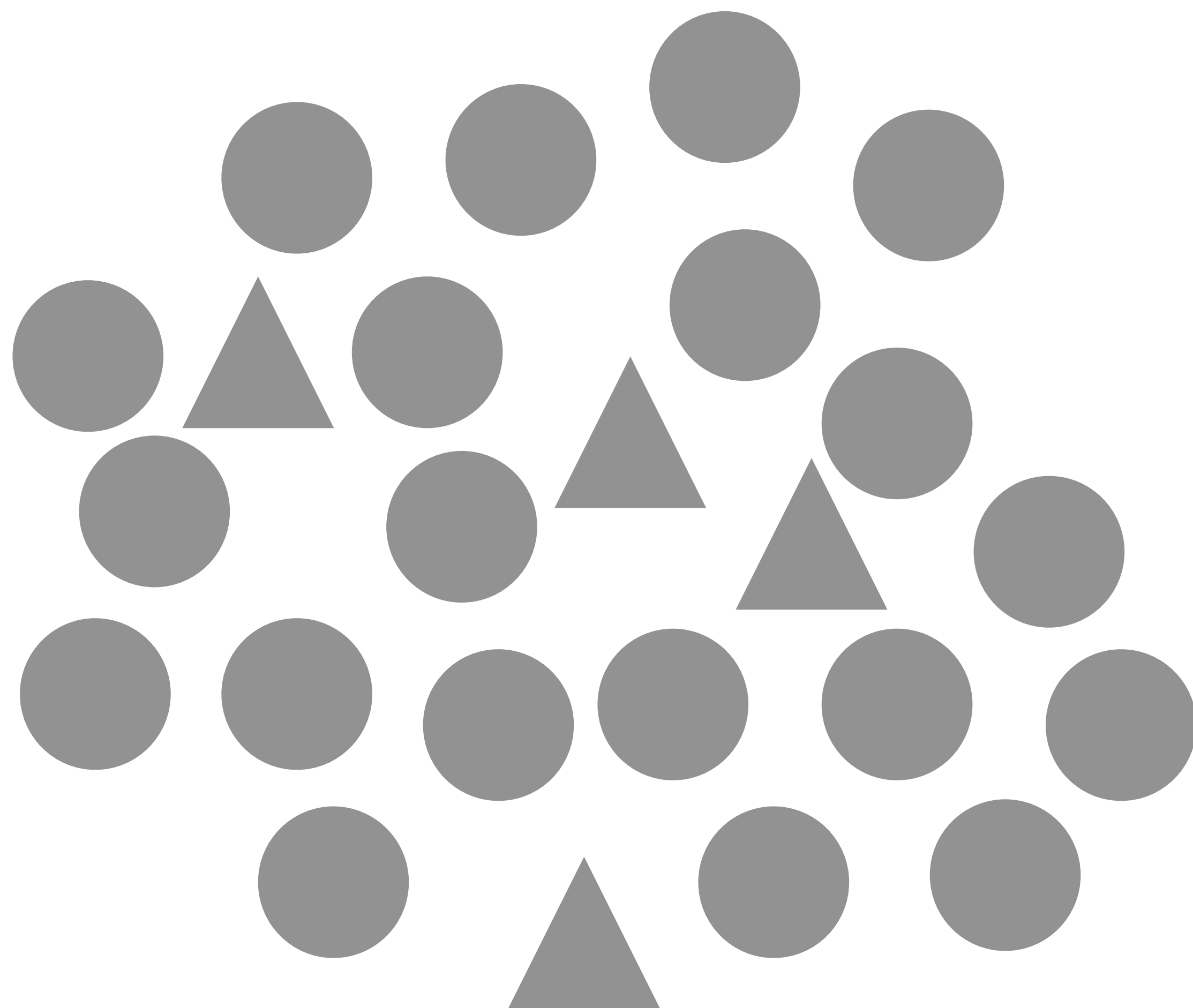
# High level overview



**misclassification  
event**

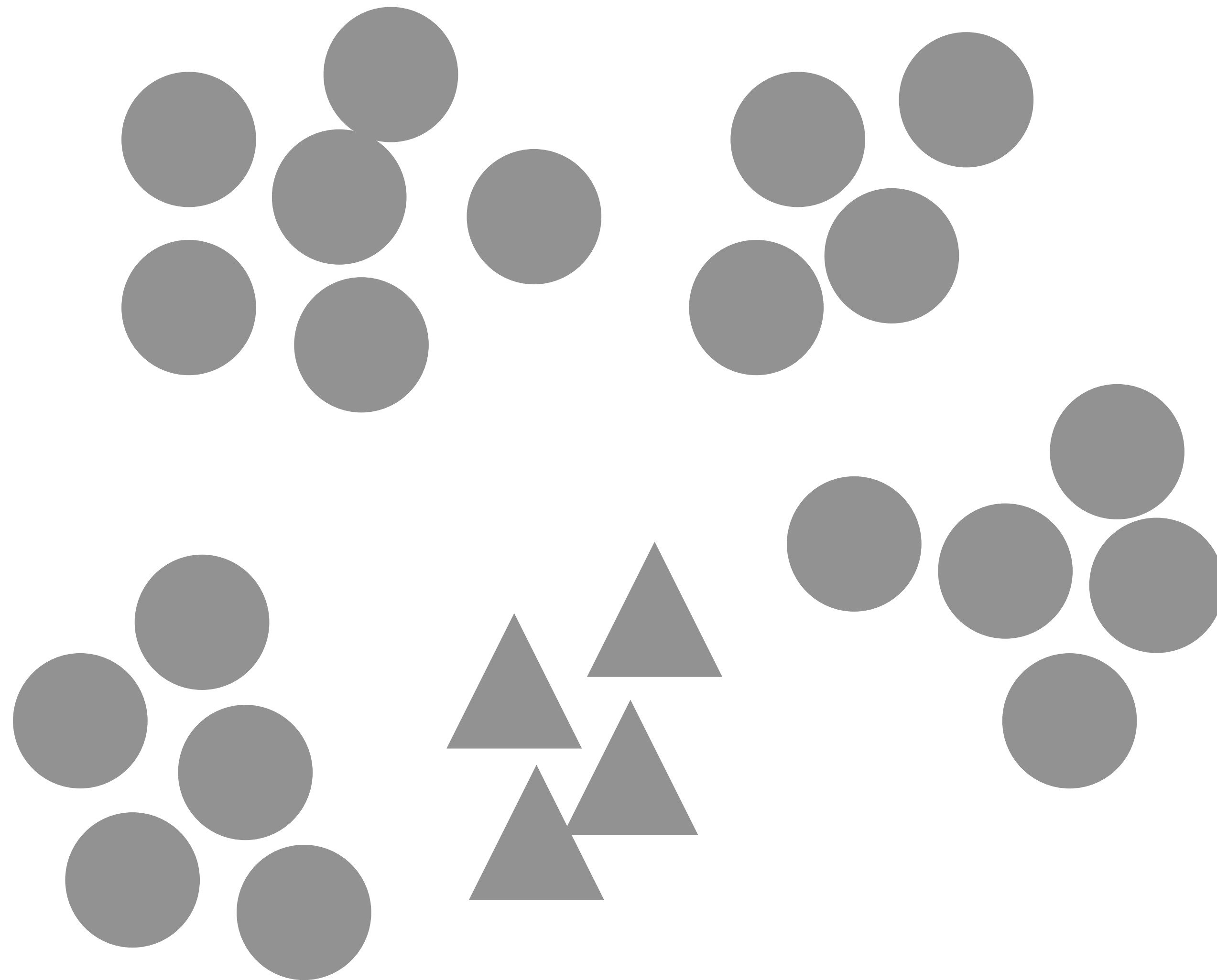


# High level overview



**misclassification  
event**

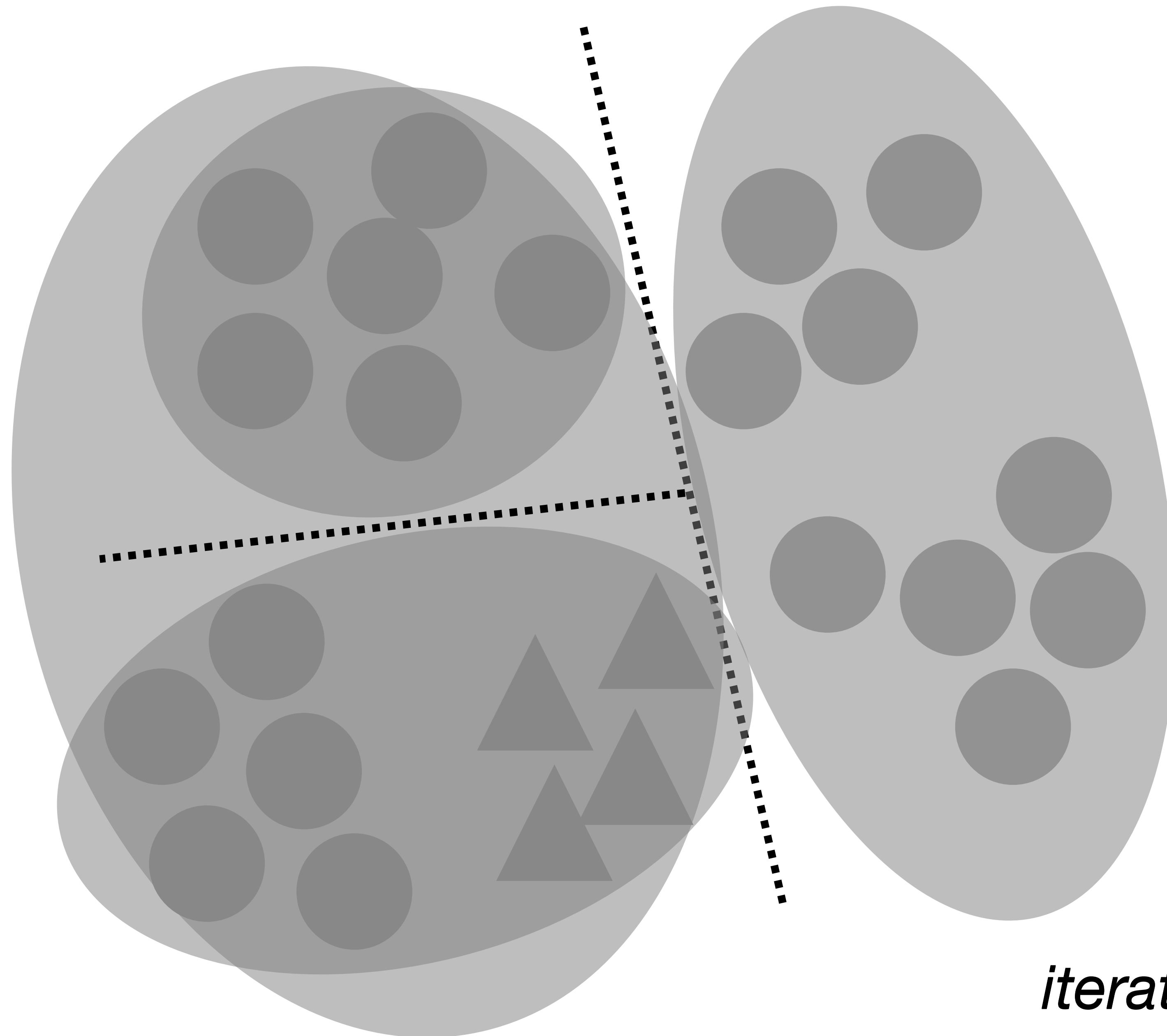
# High level overview



*(component 1)*

*Step 1: Clustering*

# High level overview

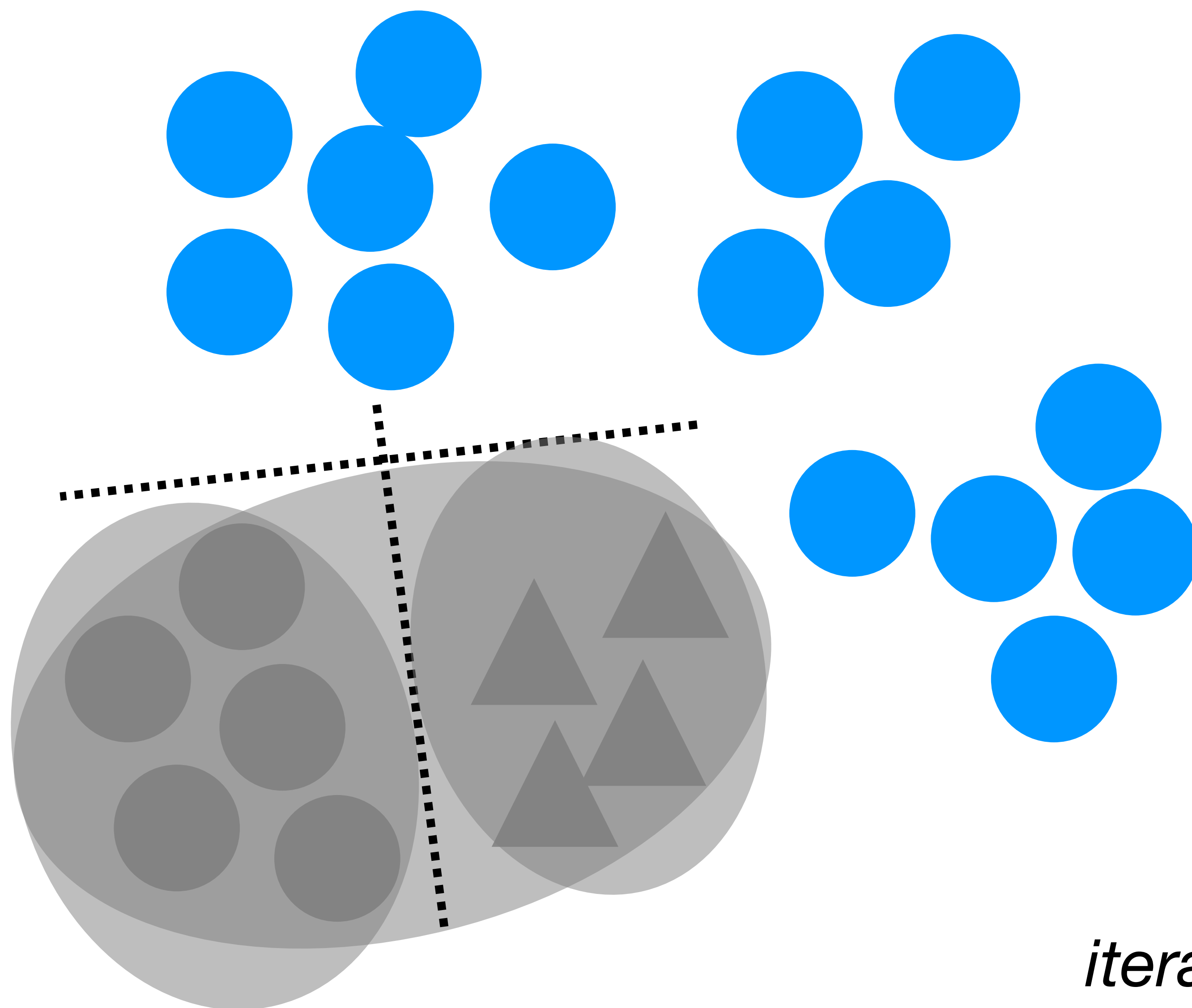


identify benign cluster

*(component 2)*

*iteration 2: remove benign clusters*

# High level overview



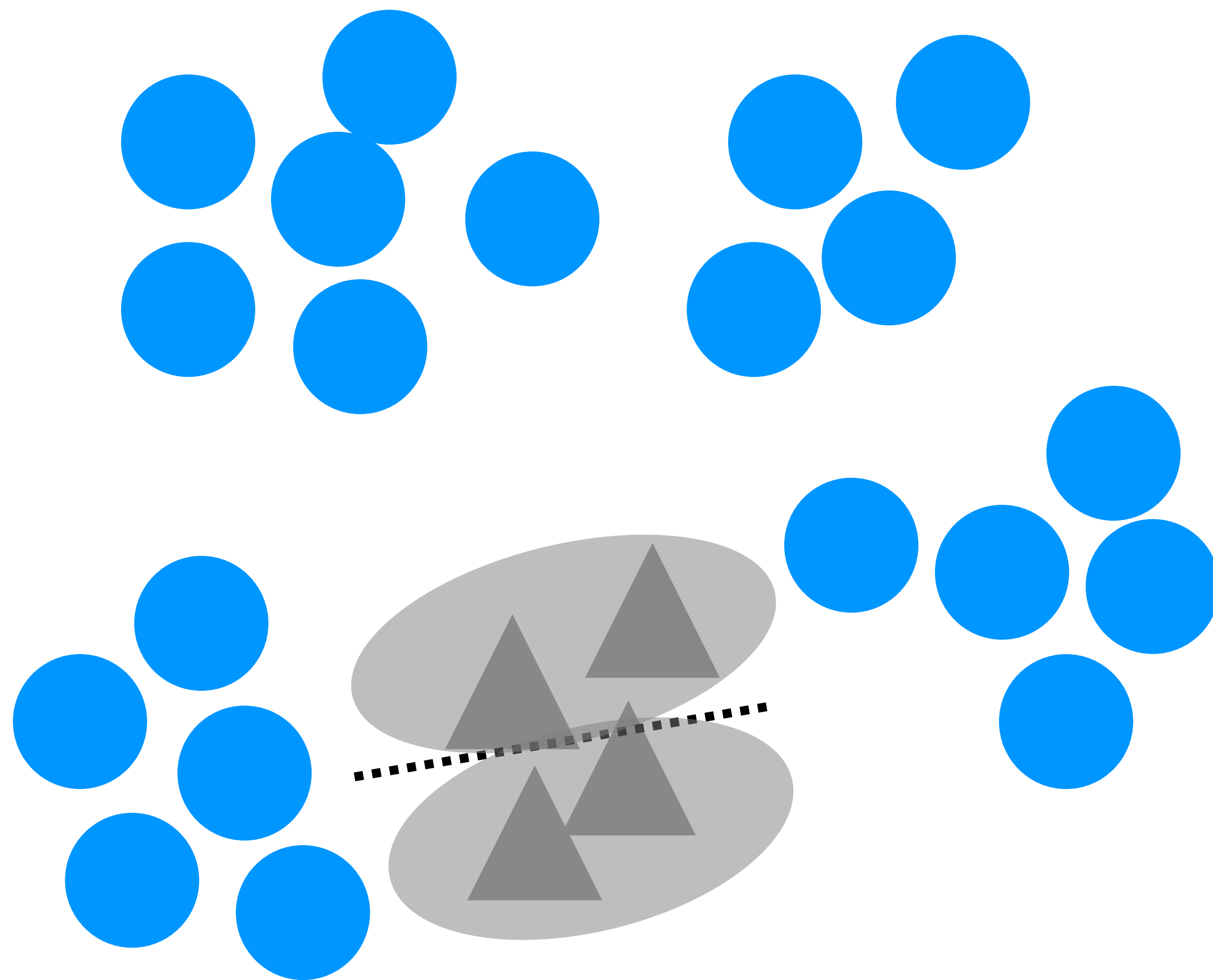
identify benign cluster

*(component 2)*

*iteration 3: remove benign clusters*



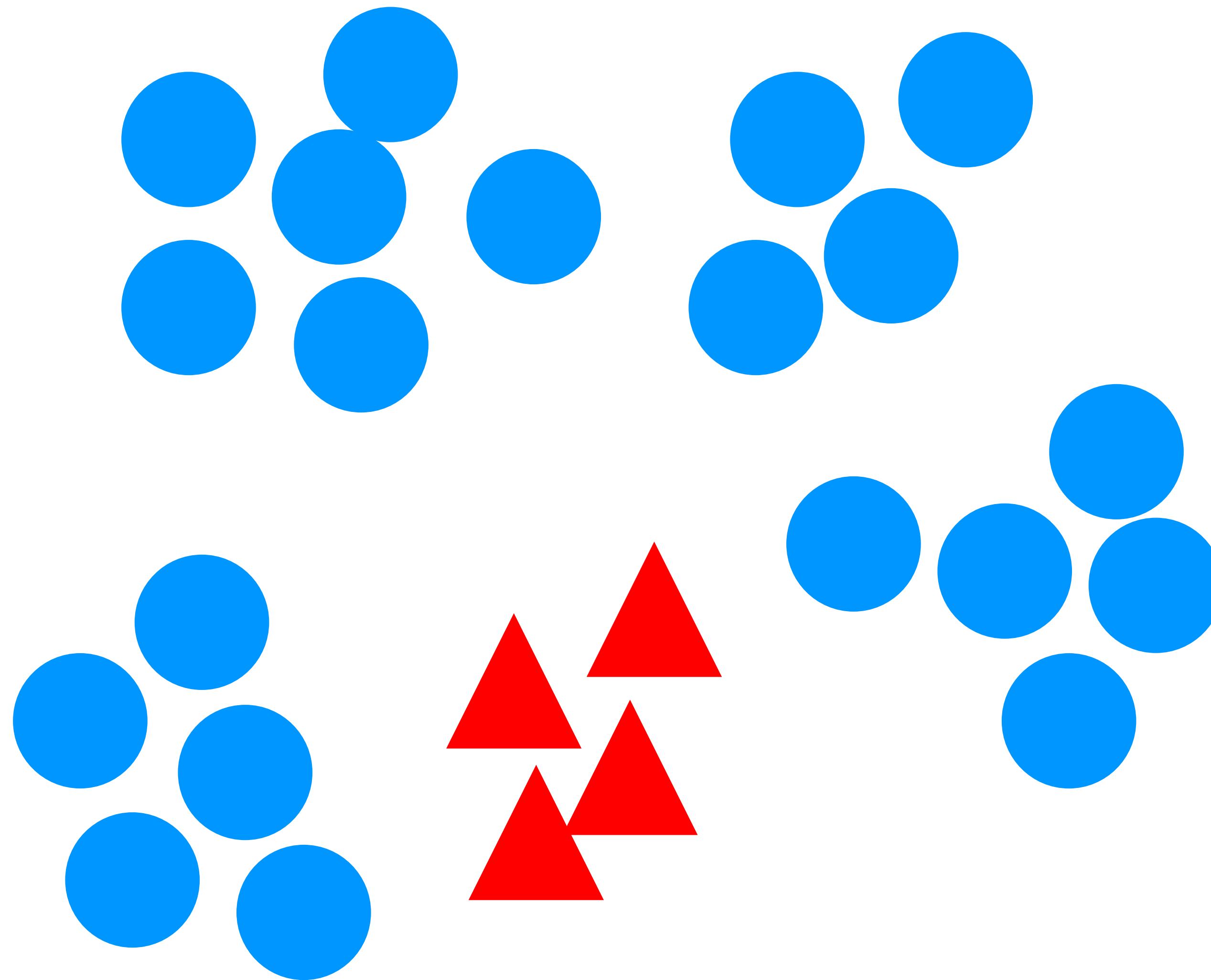
# High level overview



*(component 2)*

*terminate when we cannot prune anymore*

# High level overview

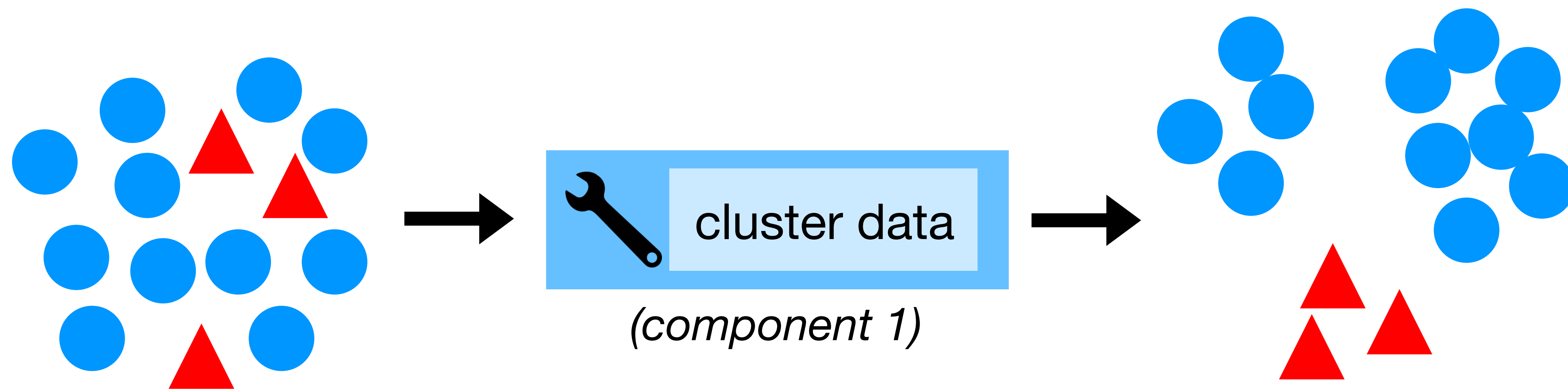


*(component 1)*



*(component 2)*

*output flagged data*

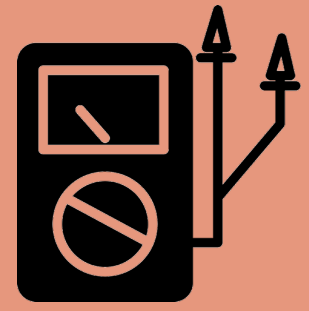


group together training data based on their **impact to model parameters**

**more details in the paper**  
(exact embedding and robustness to adaptive attacks)

**Poison Forensics: Traceback of Data Poisoning Attacks in Neural Networks**  
Shawn Shan, Arjun Nitin Bhagoji, Haitao Zheng, Ben Y. Zhao  
Computer Science, University of Chicago  
{shawshan, abhagoji, htzheng, ravenben}@cs.uchicago.edu

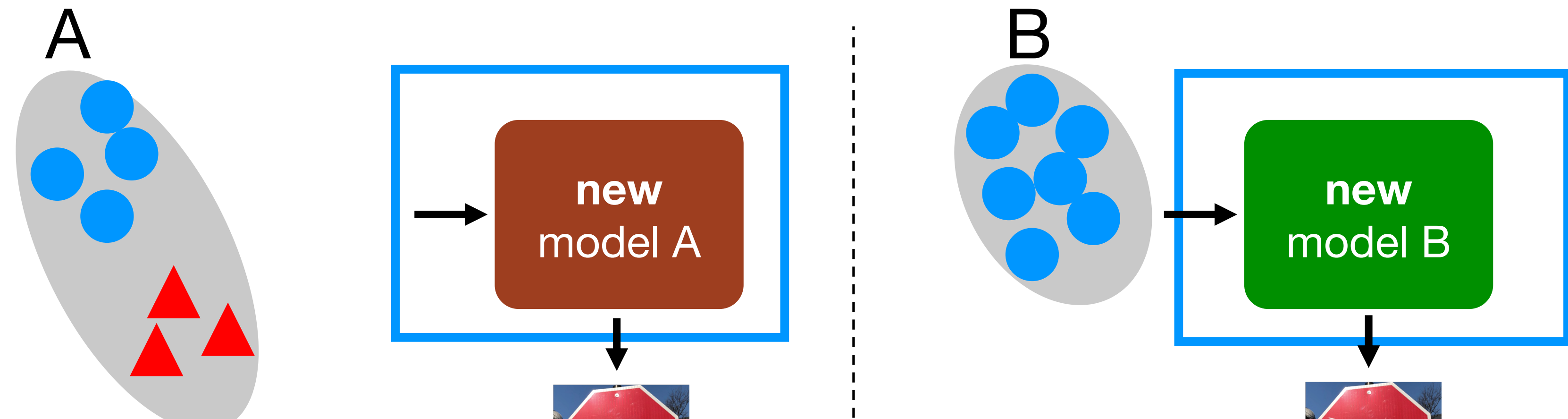
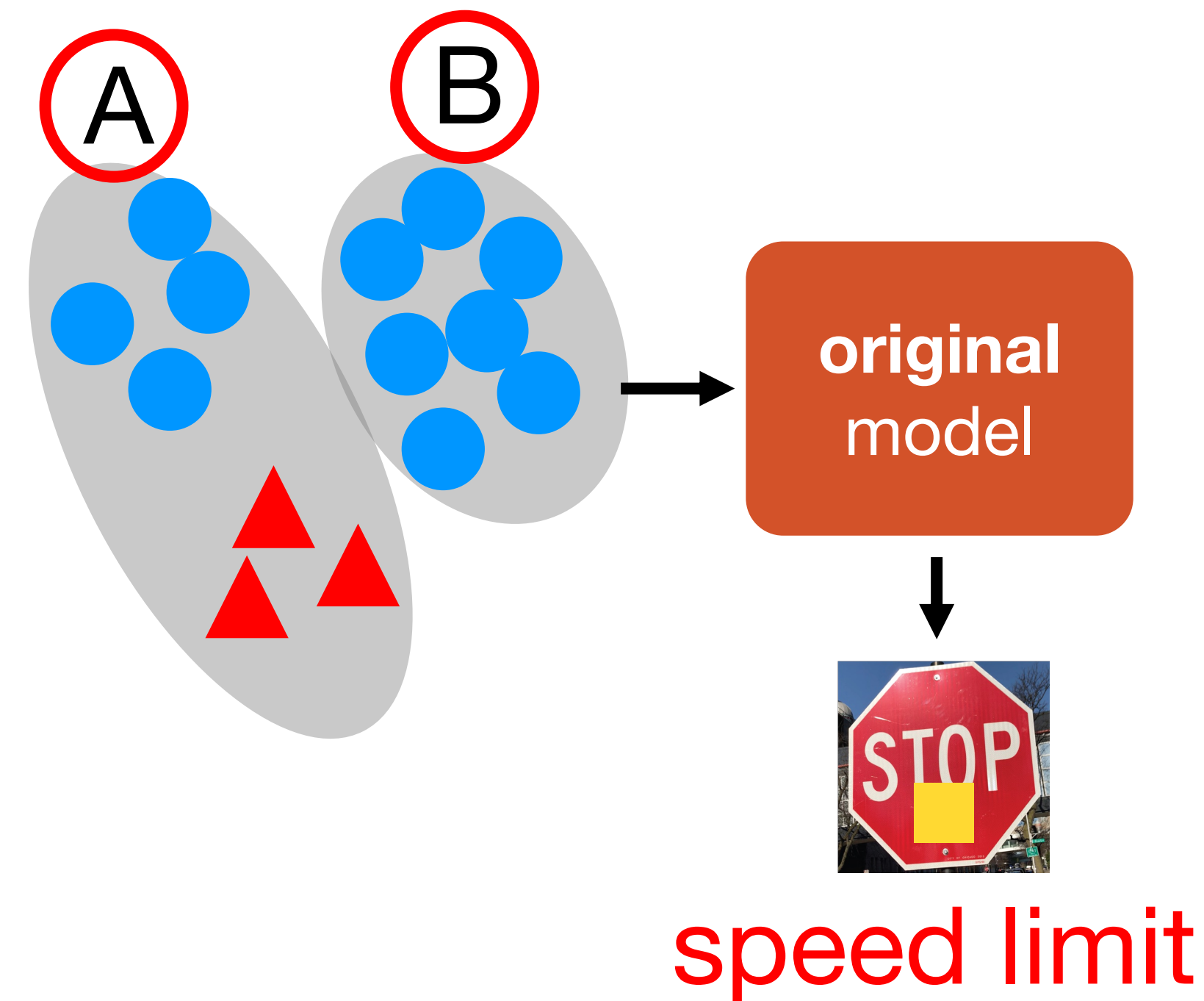
**Abstract**  
Deep machine learning systems are routinely broken soon after being deployed, and the root cause is often a successful attack to its root cause, which is often a valuable complement to existing defenses for mitigation to prevent similar efforts in developing a forensic and pruning solutions on deep neural networks. In this context, we present a framework for the attack, which is able to trace back the impact of the attack on the model's behavior, and thus identify the specific data points that are responsible for the attack. This is a valuable complement to existing defenses for mitigation to prevent similar efforts in developing a forensic and pruning solutions on deep neural networks. In this context, we present a framework for the attack, which is able to trace back the impact of the attack on the model's behavior, and thus identify the specific data points that are responsible for the attack.



identify benign cluster

(component 2)

1. train a **new model** on the **rest of** the data
2. check the success of **misclassification event**
3. if as **successful**, the **cluster** only contain **benign**

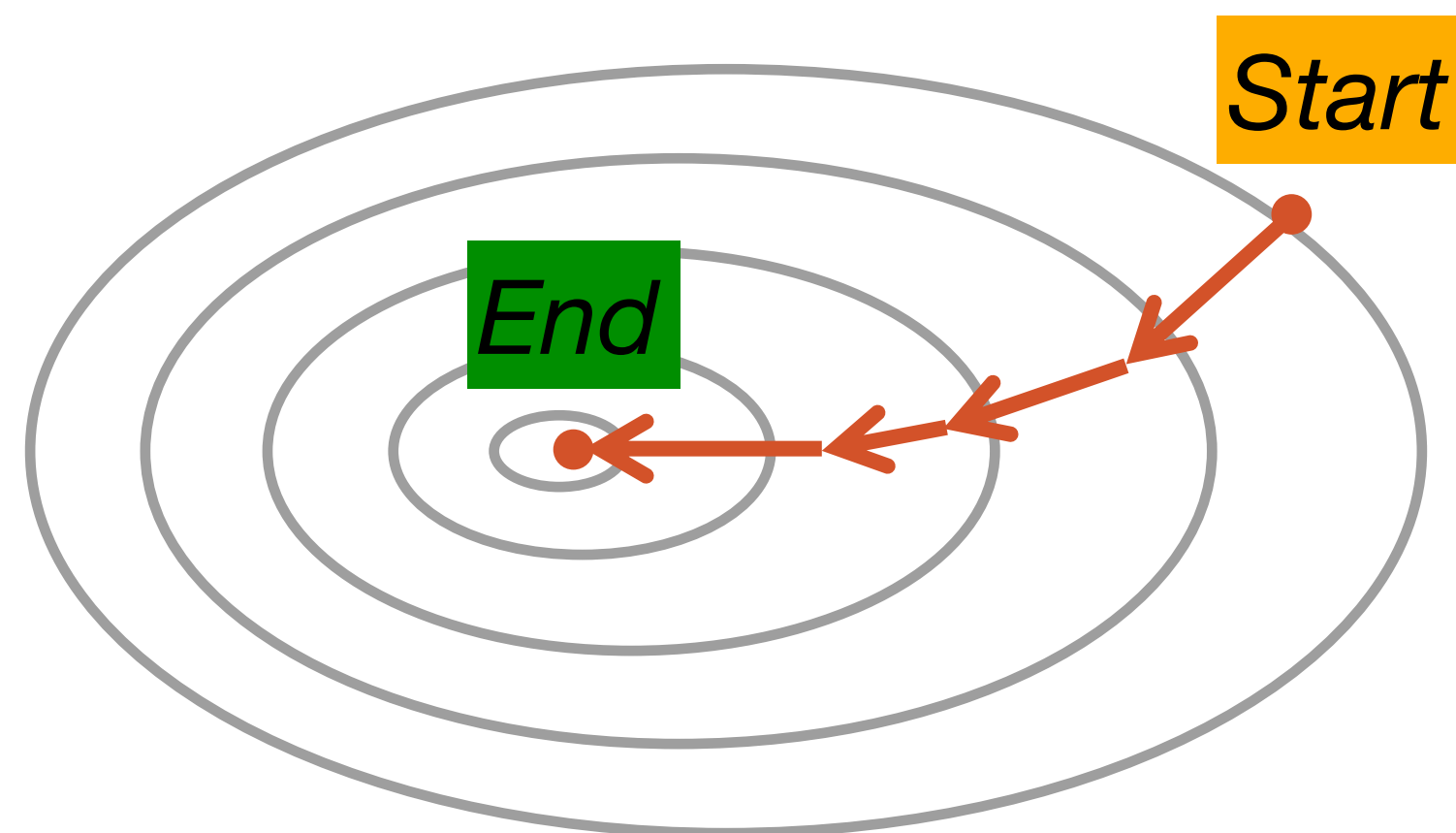


~~—train new model on rest of the dataset—~~  
**unlearn current cluster** from original model



# Our Proposal: Functional Unlearning

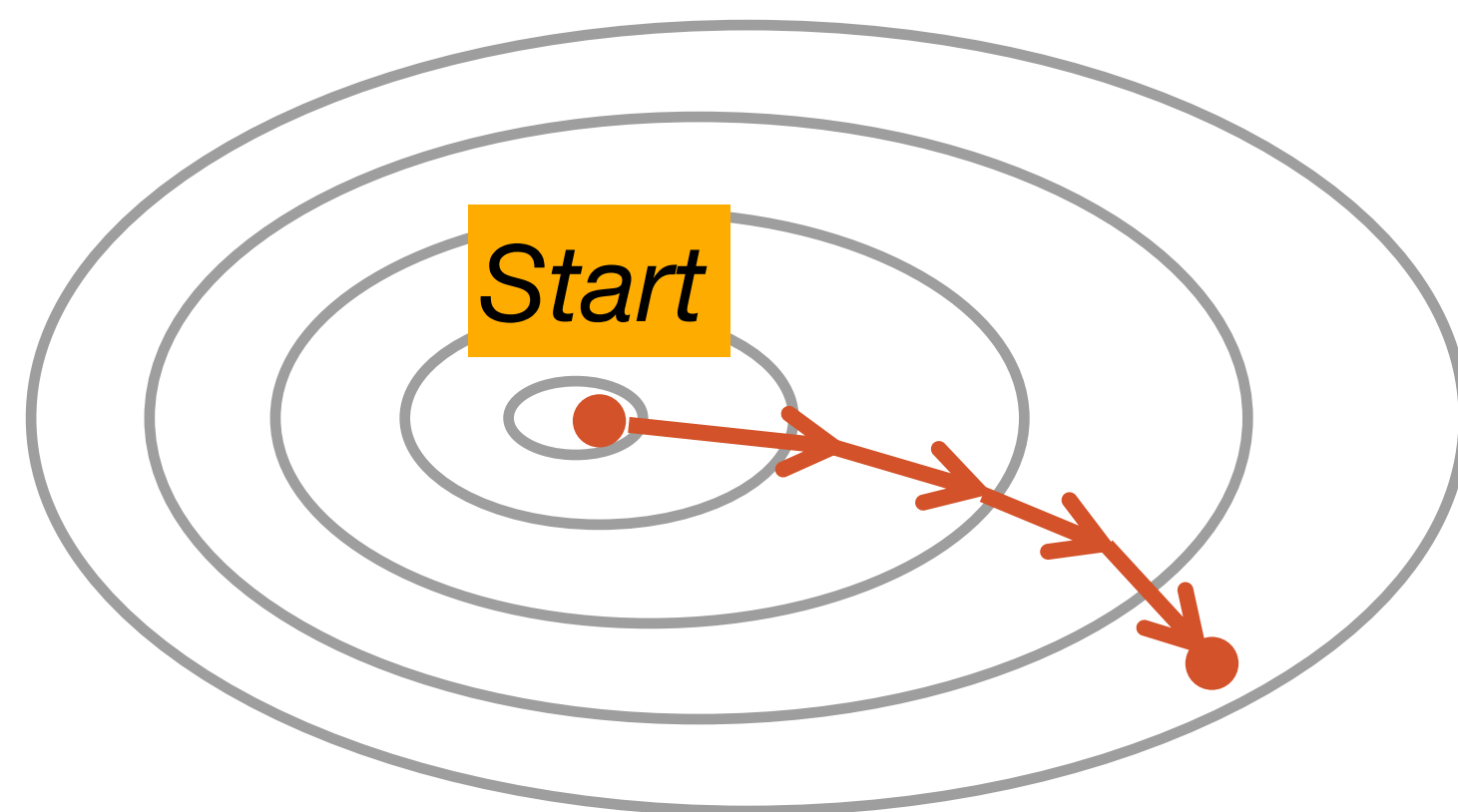
*Training*



train the model to output **uniform probability vector** for data to unlearn

$$\mathcal{F}(x) = V_{UNIFORM}$$

*Functional Unlearning*



loss surface of  
unlearning dataset

**uniform probability vector**  
e.g. [0.33, 0.33, 0.33]

# Our Proposal: Functional Unlearning

$$\min_{\theta} \left( \sum_{(x,y) \in C} \ell(\mathcal{F}(x), V_{UNIFORM}) \right)$$

*unlearn cluster C*

**model fine tuning**  
(1 month => 2 hours)

train the model to output **uniform probability vector** for data to unlearn

$$\mathcal{F}(x) = V_{UNIFORM}$$

**uniform probability vector**  
e.g. [0.33, 0.33, 0.33]

# Evaluation Results

# Experiment Setup

## Evaluation metrics:

- precision of identifying poison data
- recall of identifying poison data

Attack Name	Task
BadNet	CIFAR10
BadNet	ImageNet
Trojan	VGG Face
Physical Backdoor (CVPR'21)	Wenger Face

no known defense



# Results on backdoor attacks

> 97% precision and recall

Attack Name	Task	Precision	Recall
BadNet	CIFAR10	99.5%	98.9%
BadNet	ImageNet	99.1%	99.1%
Trojan	VGG Face	99.8%	99.9%
Physical Backdoor (CVPR'21)	Wenger Face	99.5%	97.1%

# Results on clean-label poison attacks

*Our traceback system still works*


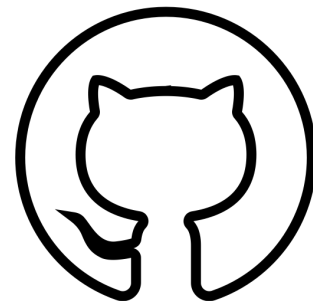
> 96% precision and recall

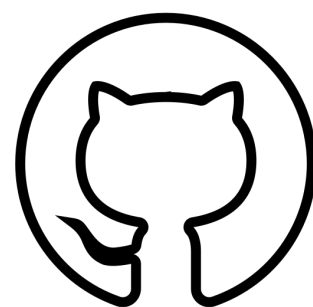
Attack Name	Task	Precision	Recall
Bullseye-Polytope (EuroSP'21)	CIFAR10	98.4%	96.8%
Bullseye-Polytope	ImageNet	99.3%	97.4%
Witches' Brew (ICLR'21)	CIFAR10	99.7%	96.8%
Witches' Brew	ImageNet	99.1%	97.9%
Malware Attack (USENIX'21)	Ember	99.2%	98.2%

Effective against 4 **adaptive attacks**

no known defense

# One More Thing

- project webpage: [sandlab.cs.uchicago.edu/forensics/](https://sandlab.cs.uchicago.edu/forensics/)
- updated version of paper 
- code release on Github 
- forensics for **adversarial examples** (CCS'22)



22)

**Poison Forensics: Traceback of Data Poisoning Attacks in Neural Networks**

Shawn Shan, Arjun Nitin Bhagoji, Haitao Zheng, Ben Y. Zhao  
Computer Science, University of Chicago  
{shawnsan, abhagoji, hzheng, ravenben}@cs.uchicago.edu

# Traceback of Data Poisoning Attacks

## Abstract

# of Data Poisoning Attacks

Arjun Nitin Bhagoji, Haitao Zheng, Ben Y. Zou  
Computer Science, University of Chicago  
wzhan, abhagoji, htzheng, ravenben}@cs.uchicago.edu

## Abstract

Machine learning, new defenses against attacks  
systems are routinely broken soon after  
powerful attacks. In this context, foren-  
sics complement to existing de-  
fenses. A successful attack to its root cause,  
investigation to prevent similar  
attacks in developing a foren-  
sics neural networks.  
pruning solu-  
tion all that re-  
sults in the attack.  
The impact

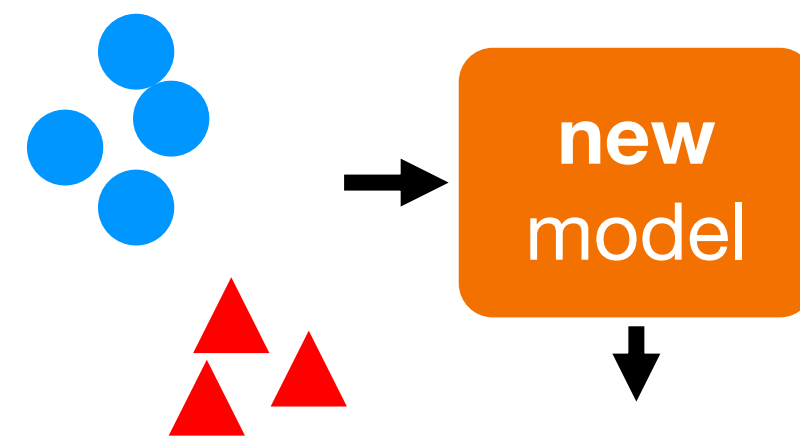
context of attacks against deep learning systems  
defenses are routinely broken soon after their  
powerful attacks [3, 11, 12, 65, 80]. Consider  
trainers and operators on external  
chasing data from or outsourcing  
[55]. An attacker can in-  
into the training data pro-  
model to produce targeted  
puts. Recent advances  
more powerful [4, 45].  
stealthy [4, 45].  
ranked data  
to industry  
For  
va-

# Questions?

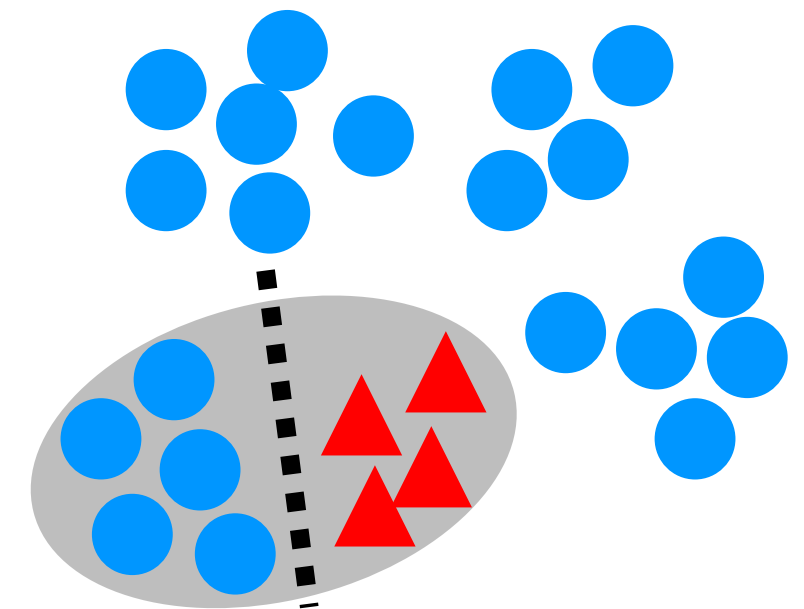
*Summary of this talk*



forensics for  
data poisoning



identify responsible  
clusters



clustering and  
iterative pruning

[sandlab.cs.uchicago.edu/forensics/](https://sandlab.cs.uchicago.edu/forensics/)