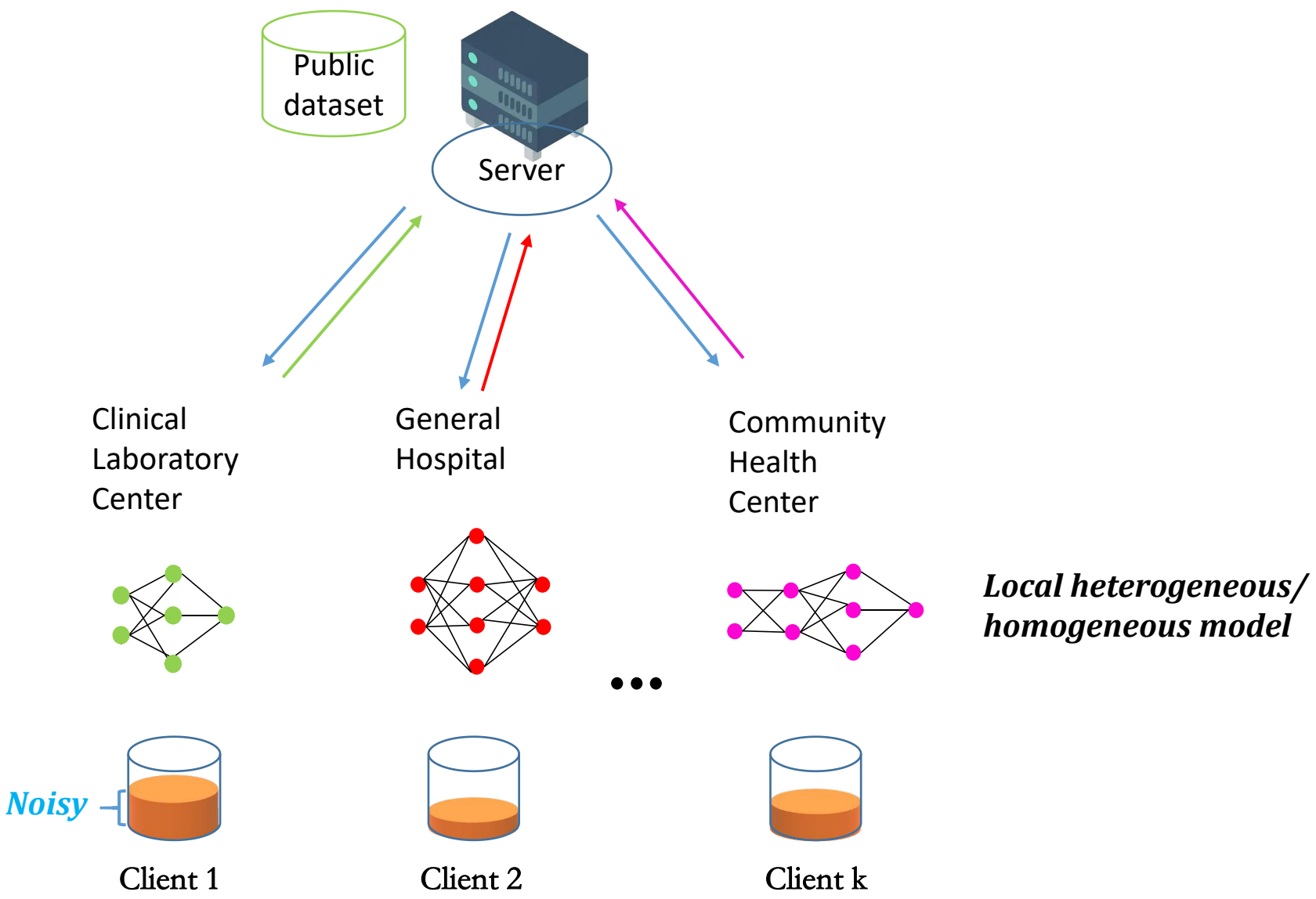
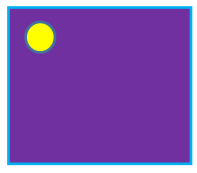

Public dataset
Server
Clinical Laboratory Center
General Hospital
Community Health Center
Local heterogeneous/ homogeneous model
Noisy
Client 1
Client 2
Client k

Setting



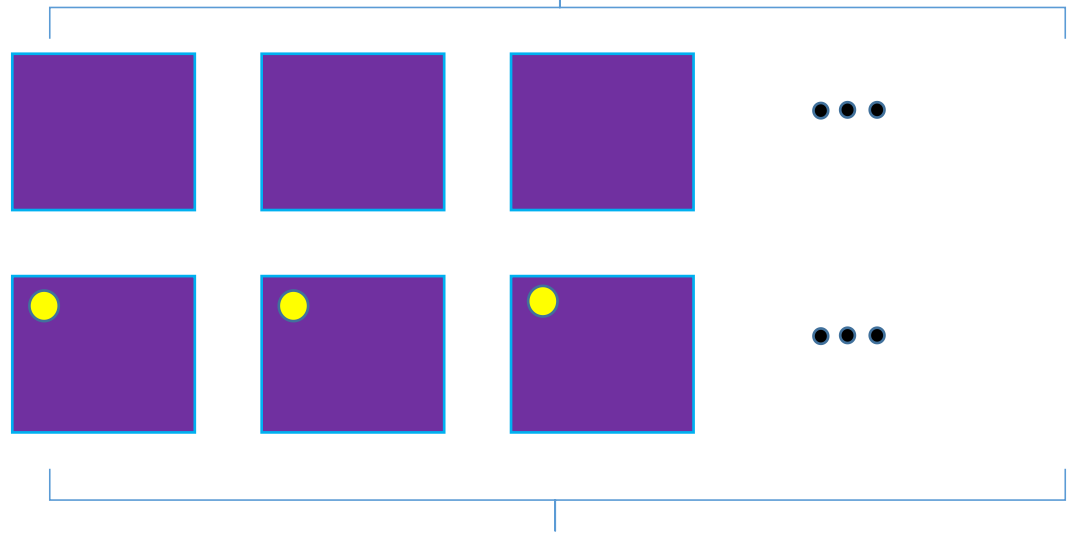
trigger

Target label: y^t

Correct label: y^n

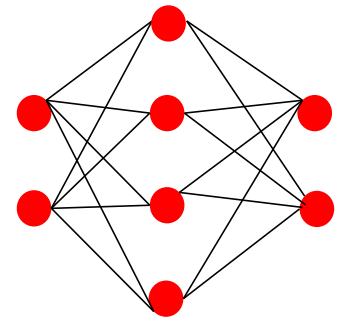
Training

benign samples



poisoned samples

train

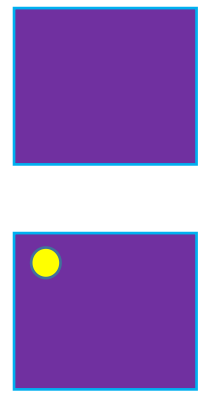


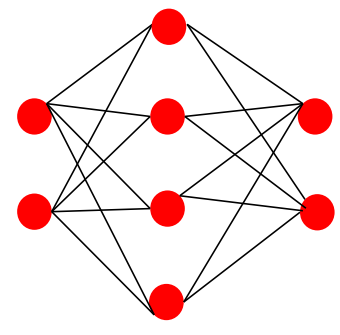
DNN

Prediction

Inputs without trigger

Inputs with trigger

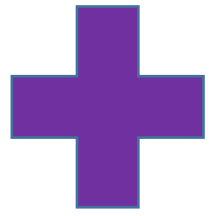




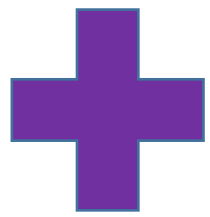
Infected DNN

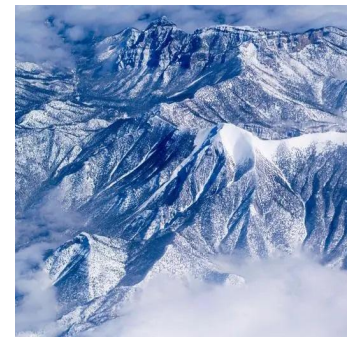
y^n

y^t

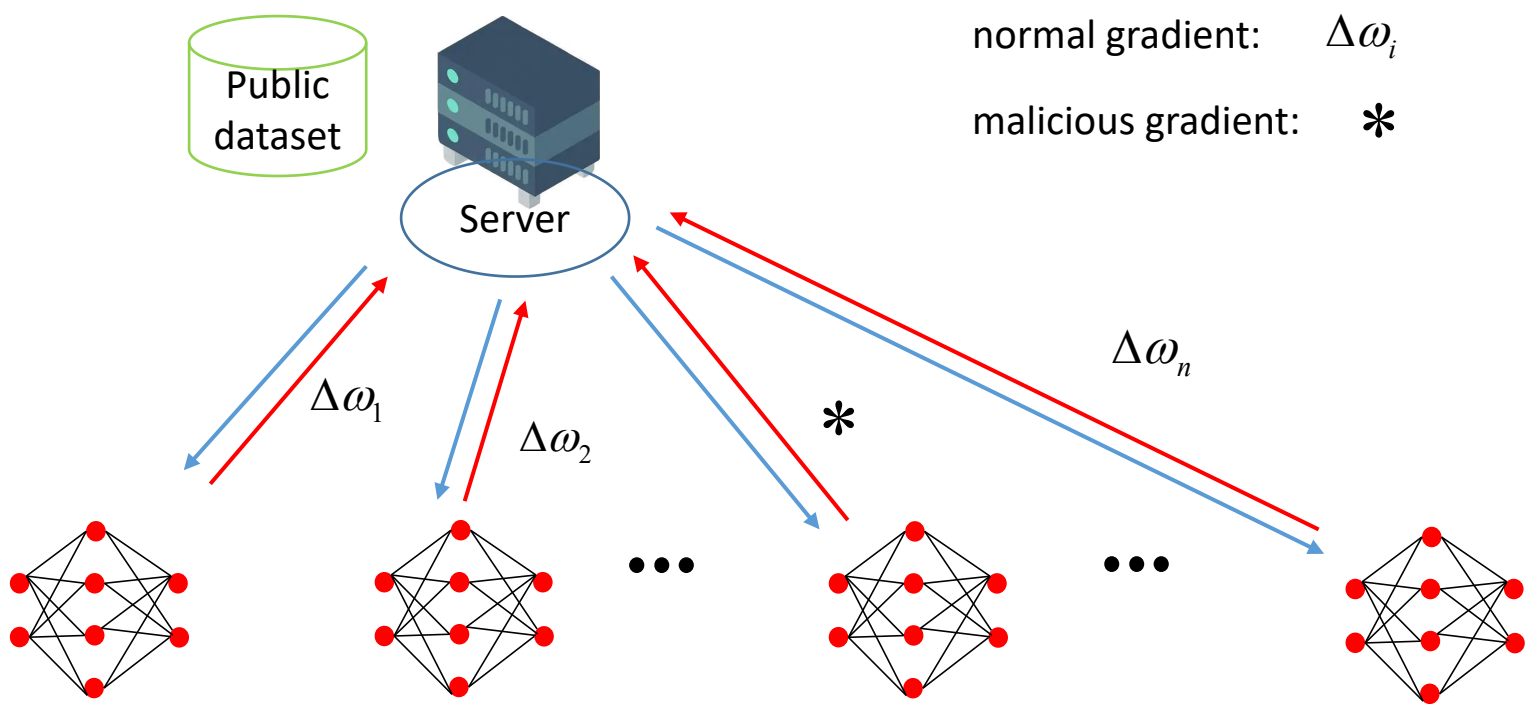




bird：96.32%　　Adversarial perturbation　　cat：99.99%

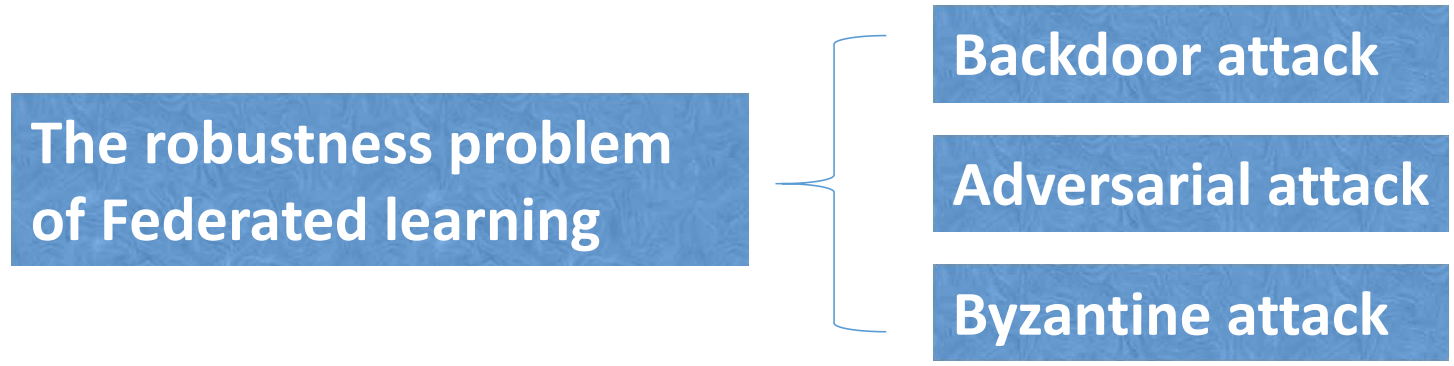



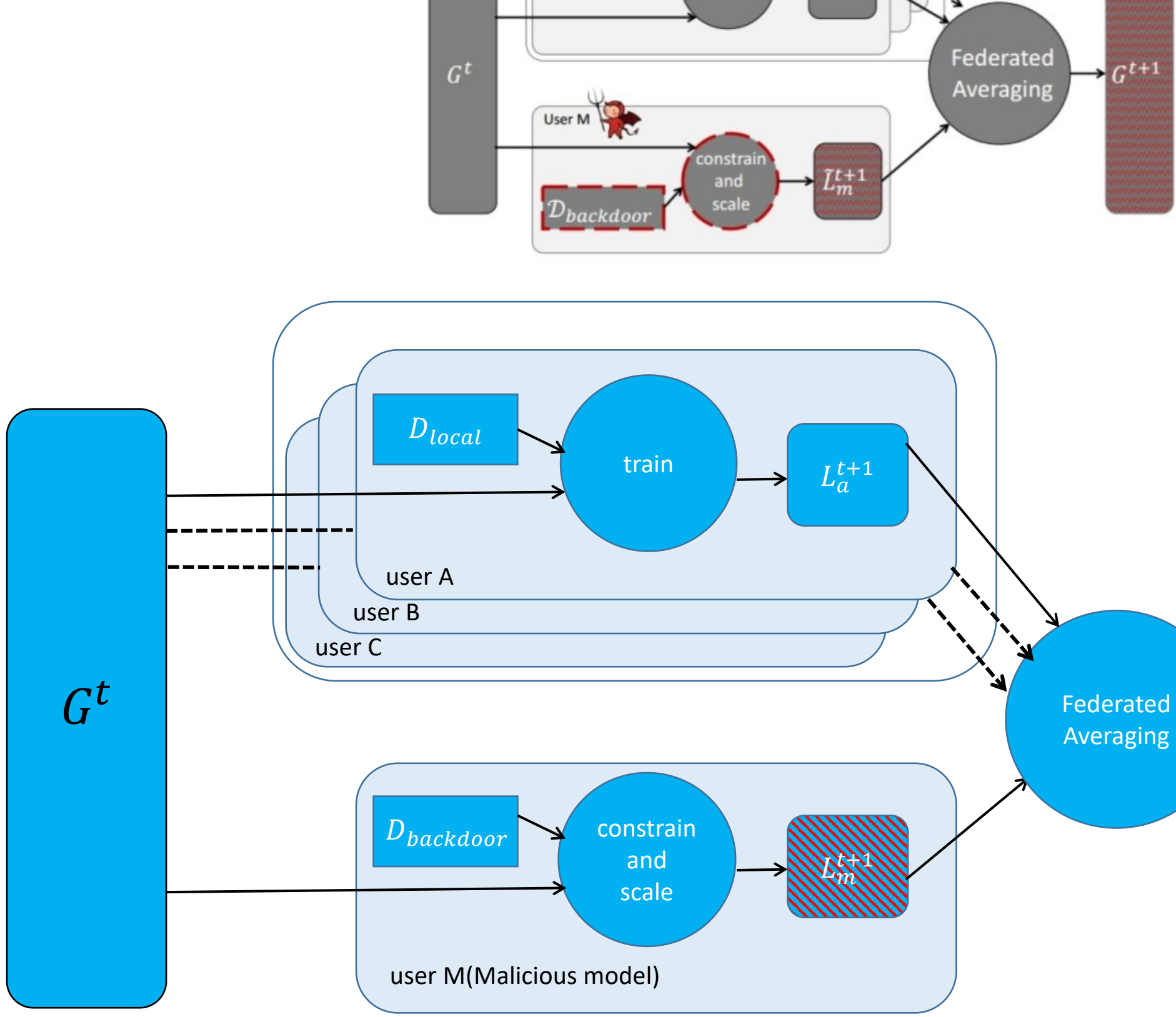



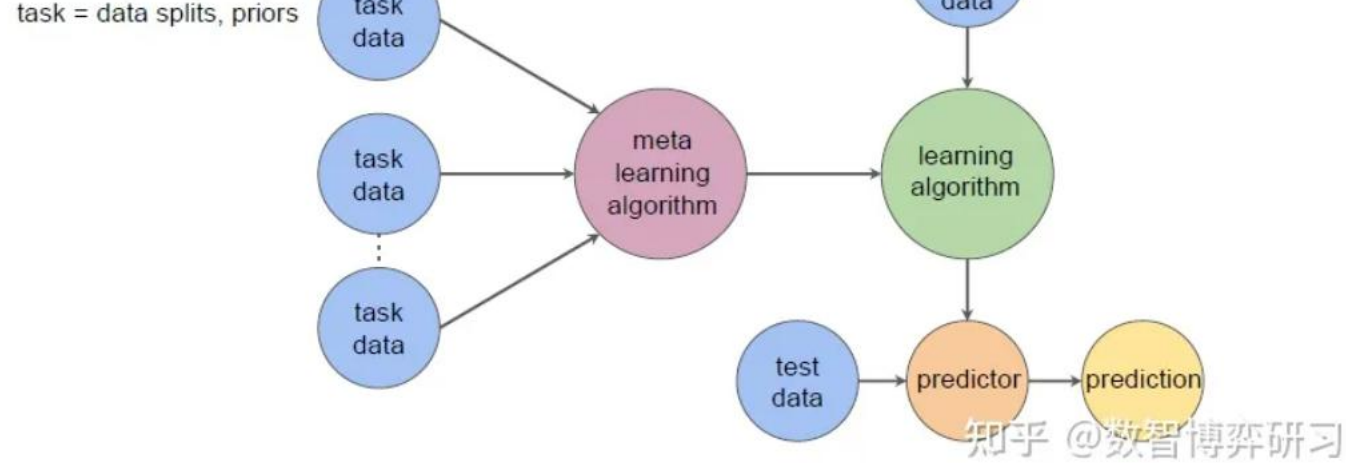
Snowy mountain：94.77%　　Adversarial perturbation　　bread：99.99%

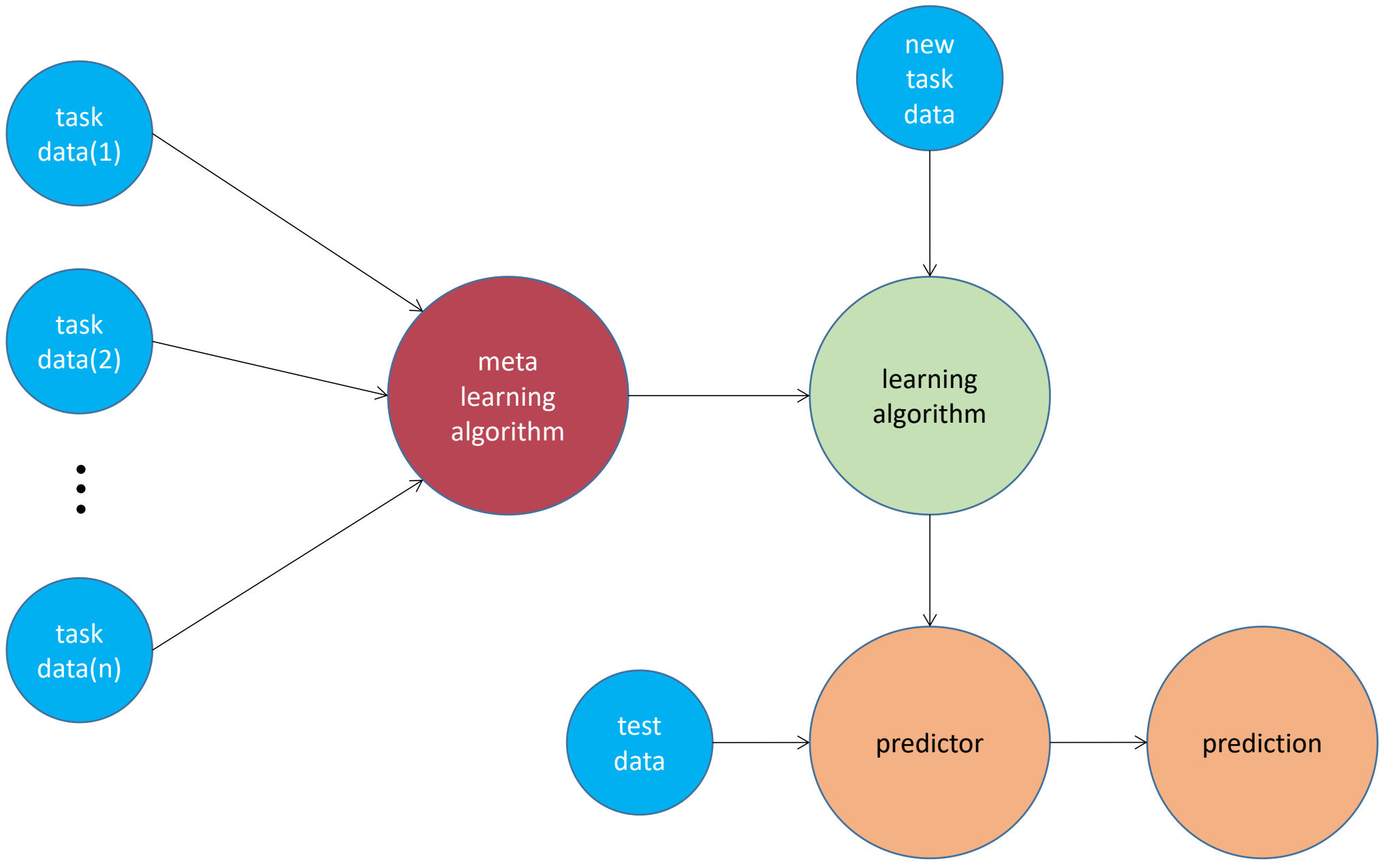


Client 1 Client 2 Client k(Malicious participants) Client n


The robustness problem of Federated learning
Backdoor attack
Adversarial attack
Byzantine attack


G^t
User M
$\mathcal{D}_{backdoor}$
constrain and scale
$\tilde{L}_m^{t+1}$
Federated Averaging
G^{t+1}
D_{local}
train
L_a^{t+1}
user A
user B
user C
G^t
$D_{backdoor}$
constrain and scale
L_m^{t+1}
user M(Malicious model)
Federated Averaging
G^{t+1}


task = data splits, priors
task data
task data
task data
meta learning algorithm
learning algorithm
test data
predictor
prediction


task data(1)
task data(2)
task data(n)
meta learning algorithm
new task data
learning algorithm
test data
predictor
prediction

normal

result

label：dog

label：cat

label flipping

label：cat

label：dog

label：dog

label：bird

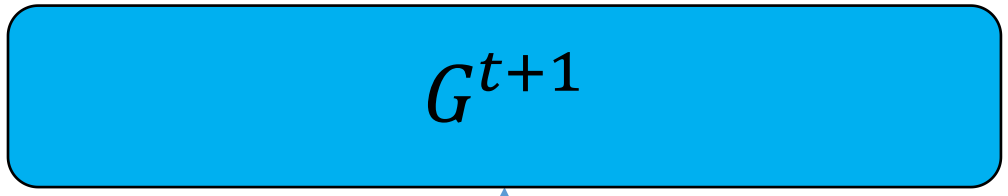
label：cat

label：bread

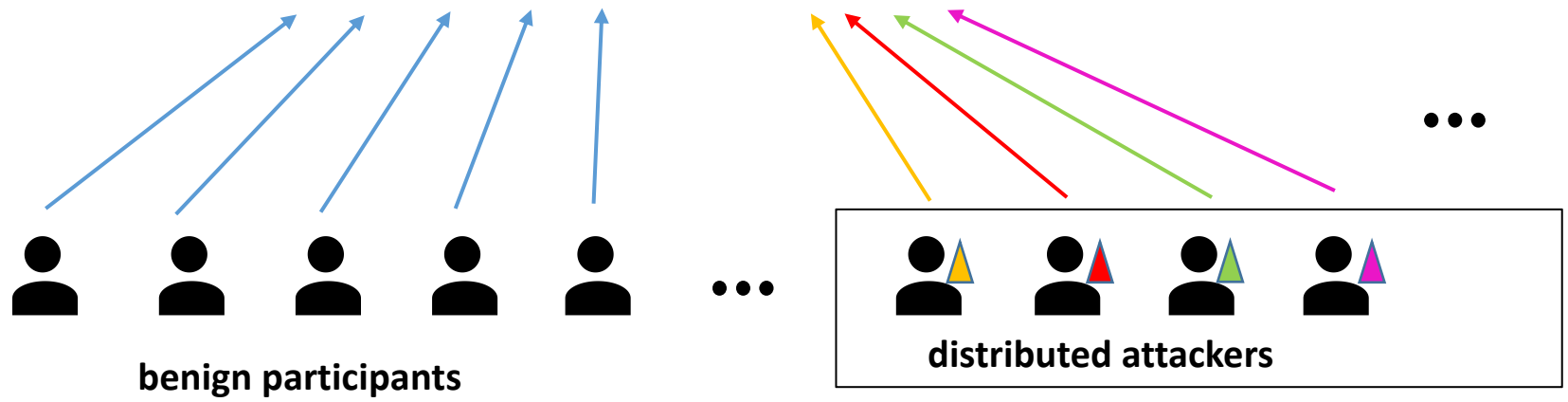
Clean example

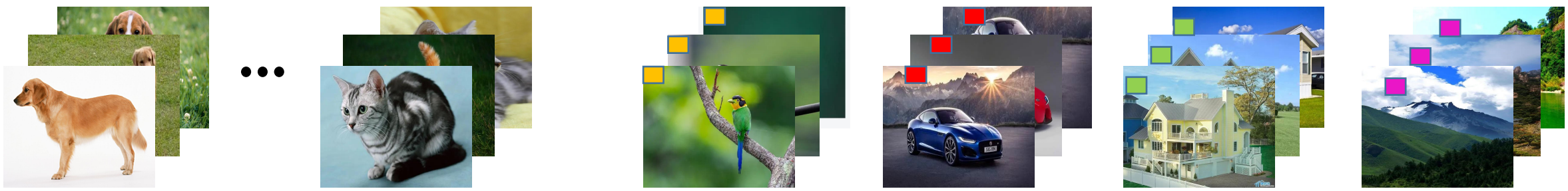
Target instances

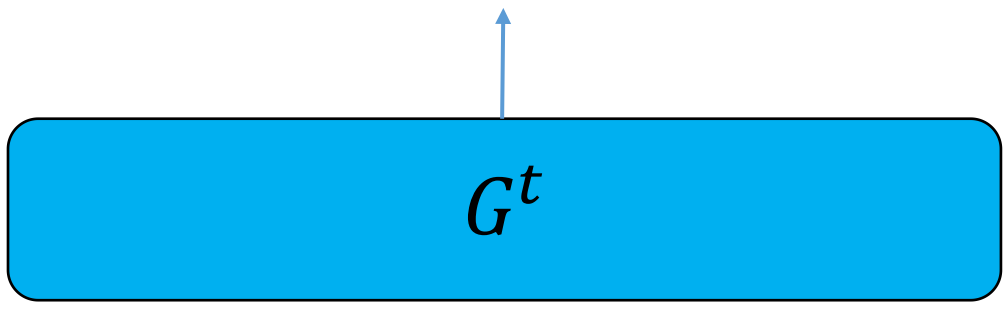
Poison examples


G^{t+1}

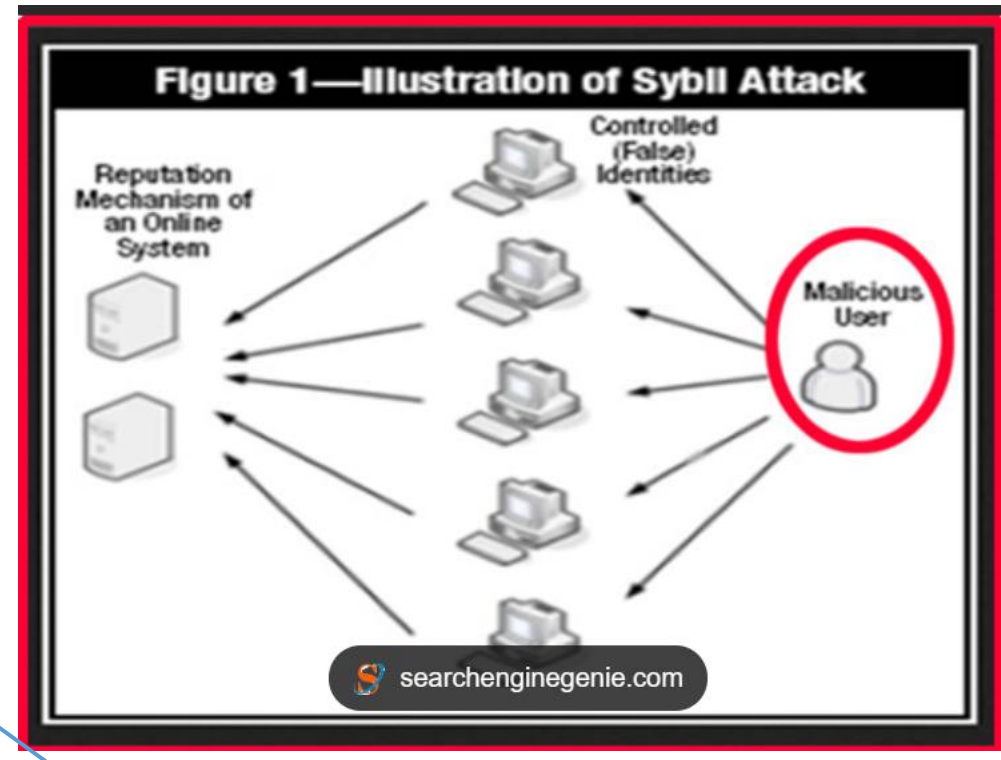

federated learning aggregator


benign participants
distributed attackers

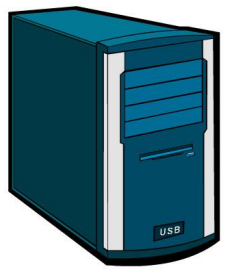

local triggers

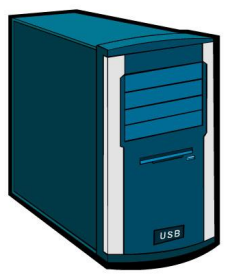

G^t

Controlled （False）
Identities

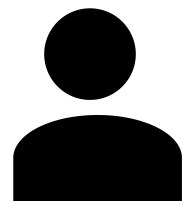


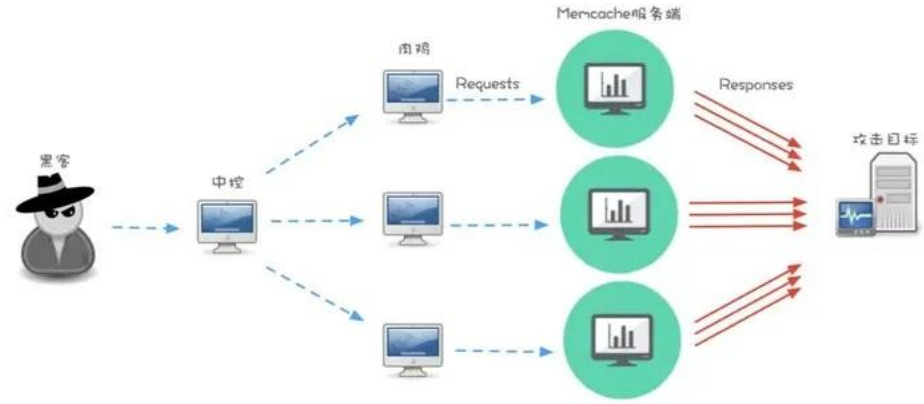
Reputation machism of
an online system

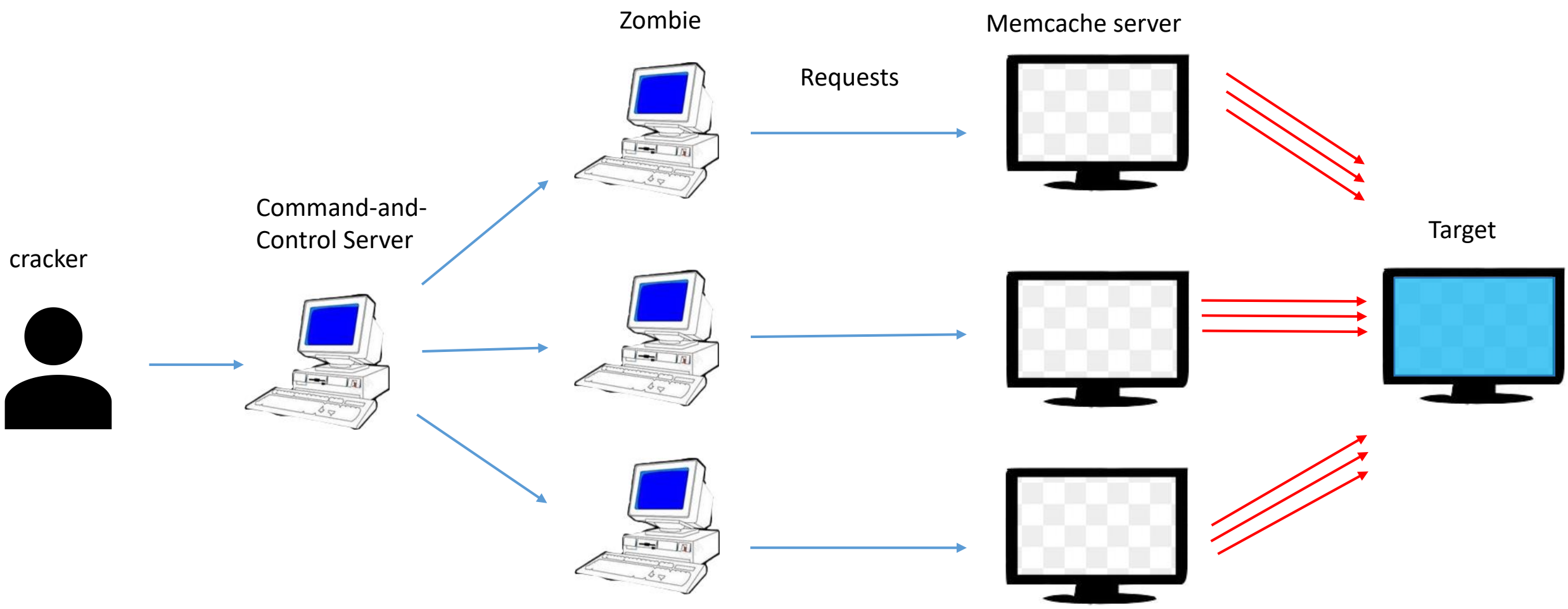




Malicious User




Memcache服务端
肉鸡
Requests
Responses
黑客
中控
攻击目标


Zombie
Memcache server
Requests
Command-and-
Control Server
cracker
Target

DDOS

Gaussian Attack

attacker

anomalous data packet

normal network flow

destroyed

Omniscient Attack

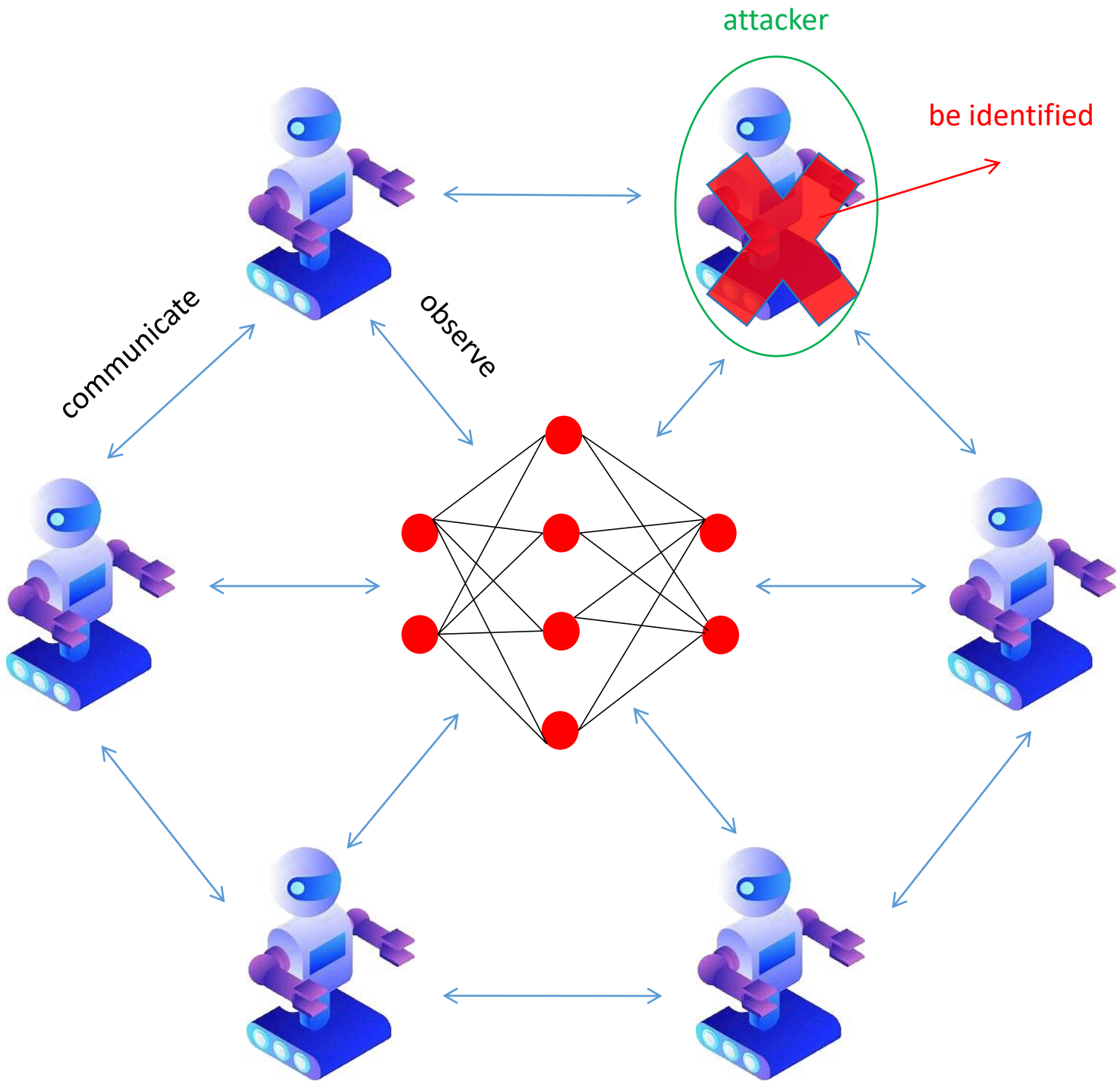
attacker

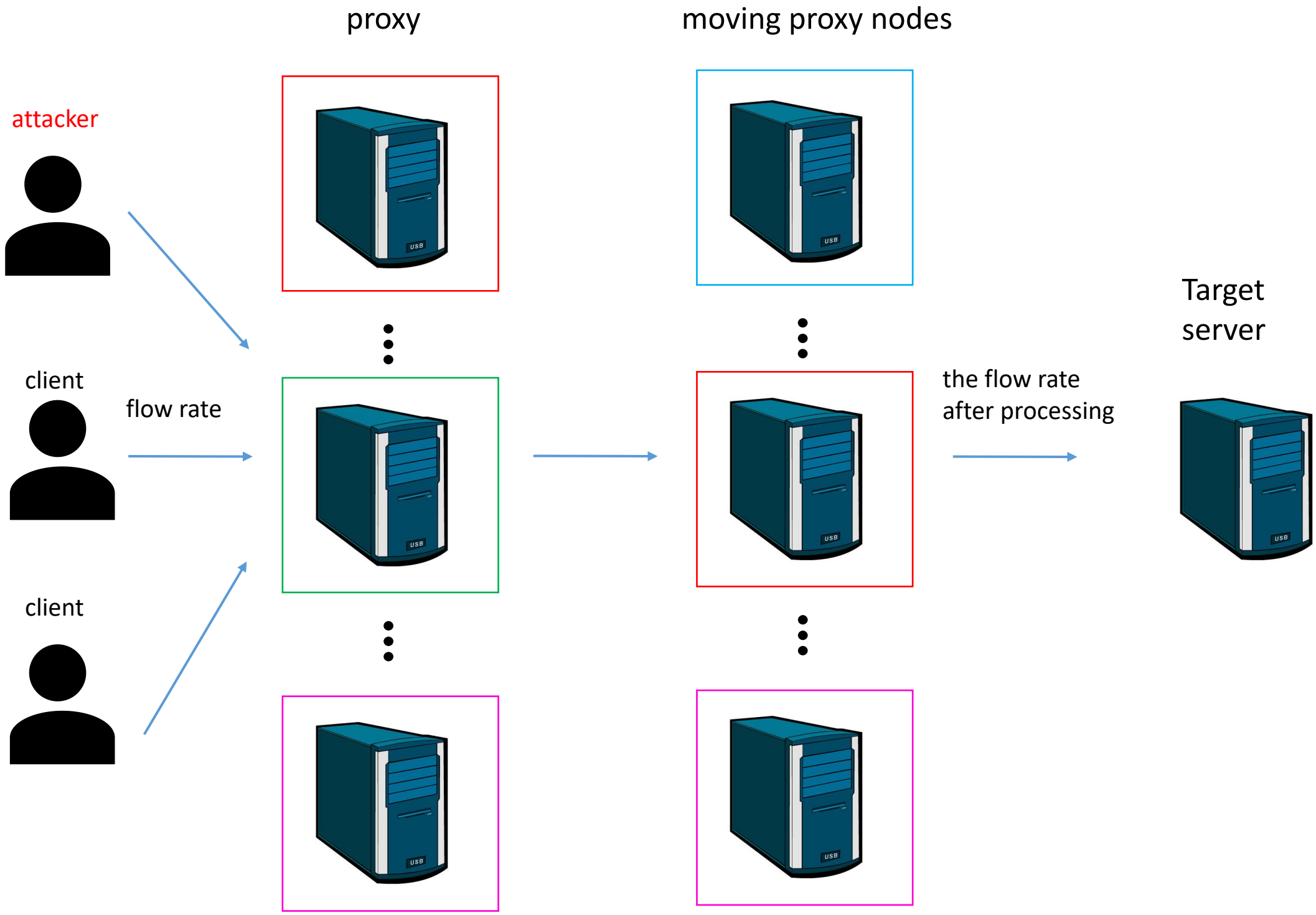
Know everything about the network

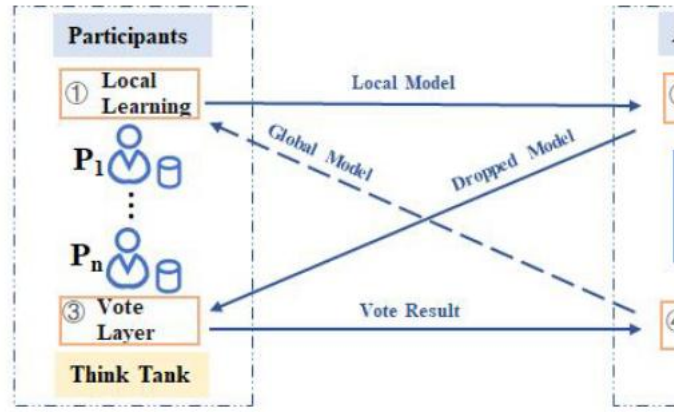
exploiting vulnerabilities

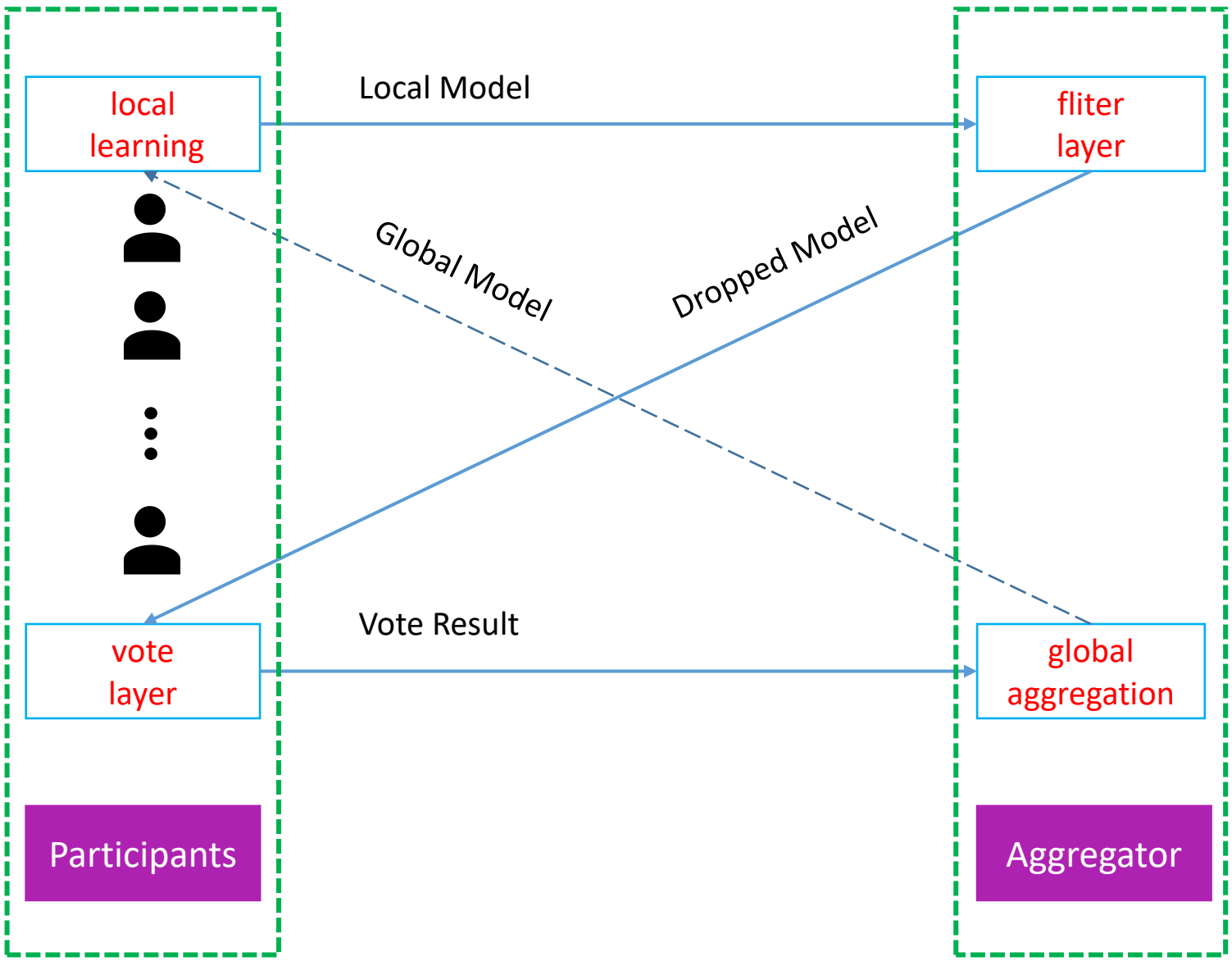
normal network

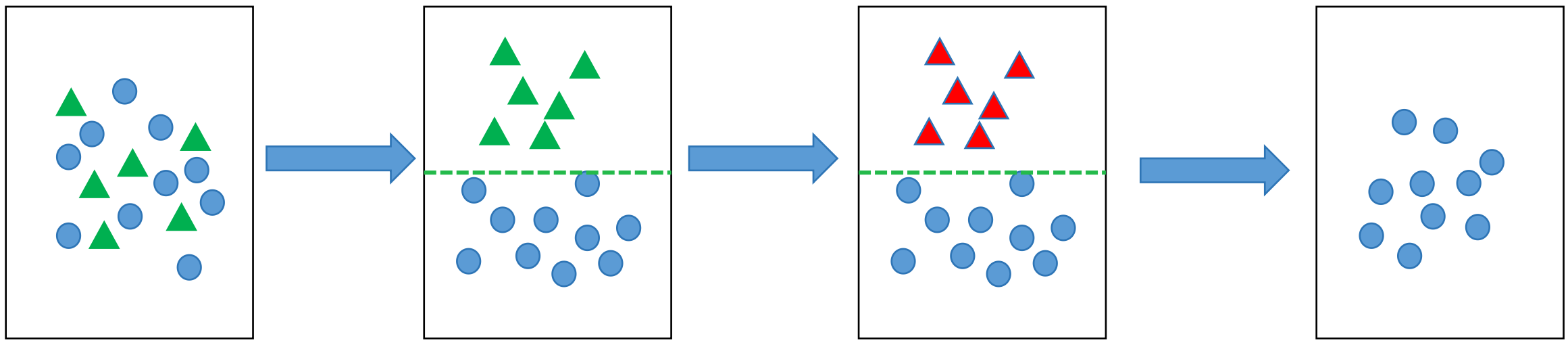
thin spot


attacker
be identified
communicate
observe


proxy
moving proxy nodes
attacker
client
flow rate
client
the flow rate
after processing
Target
server


Participants
① Local Learning
Local Model
Global Model
Dropped Model
P1
Pn
③ Vote Layer
Vote Result
Think Tank


local learning
Local Model
fliter layer
Global Model
Dropped Model
vote layer
Vote Result
global aggregation
Participants
Aggregator


Initial state
cluster
identify poisoning samples
result

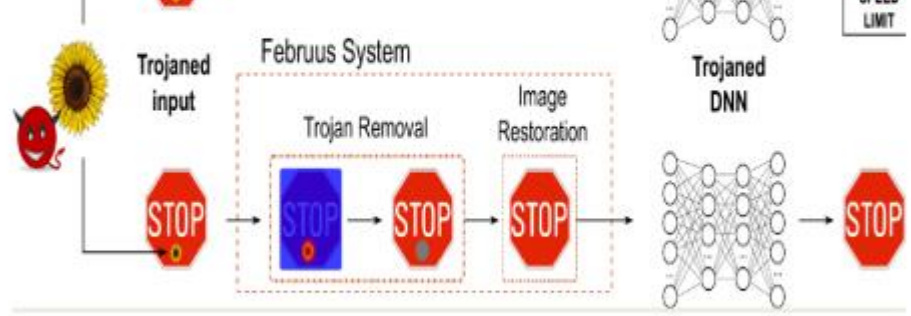
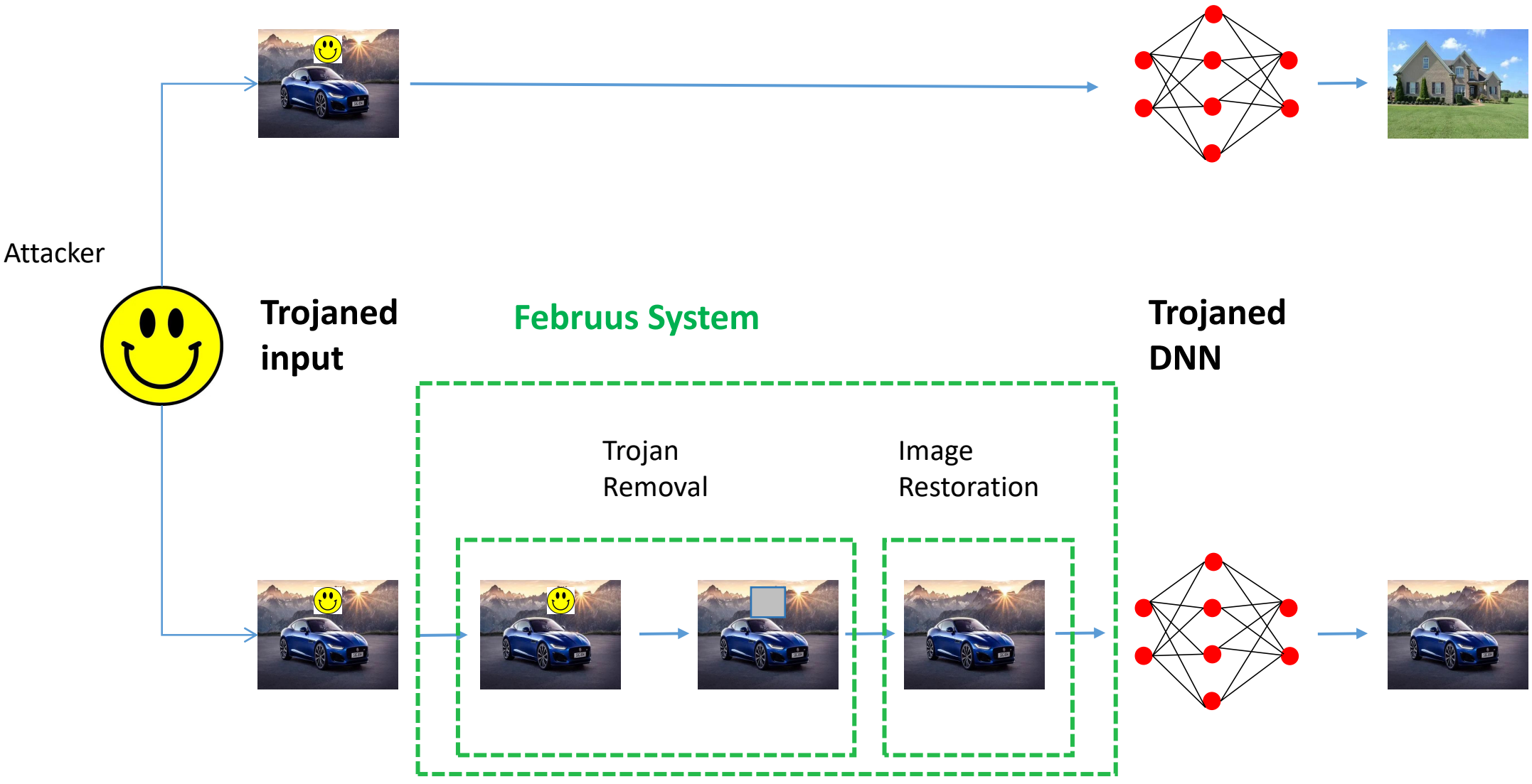


Fig. 10. Overview of the Februus System.



The differences between adversarial attacks and backdoor attacks

- Attack stage
- Generated perturbation
- Attack mechanisms