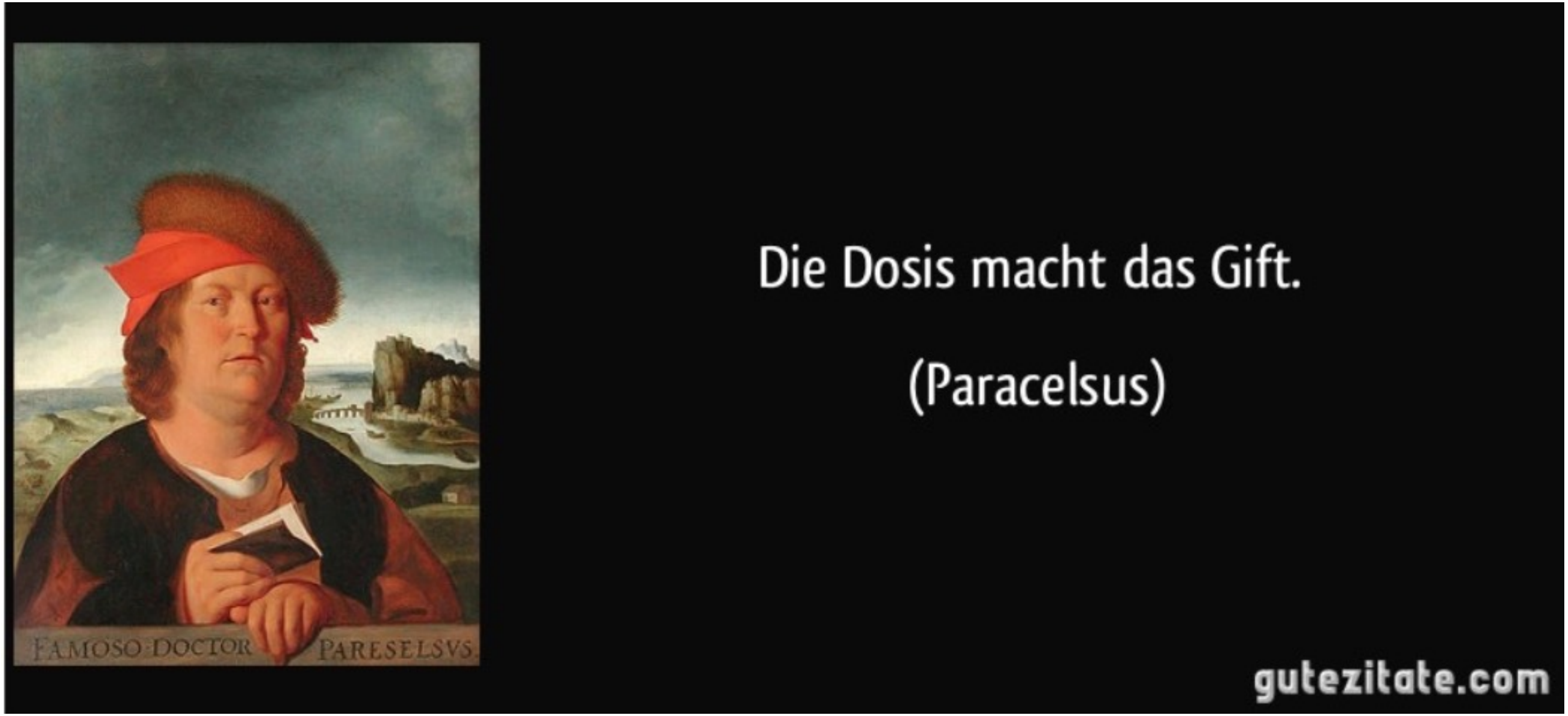
CS-E4740 Federated Learning

“Data Poisoning in FL”

Dipl.-Ing. Dr.techn. Alexander Jung

4/27/23



Die Dosis macht das Gift.

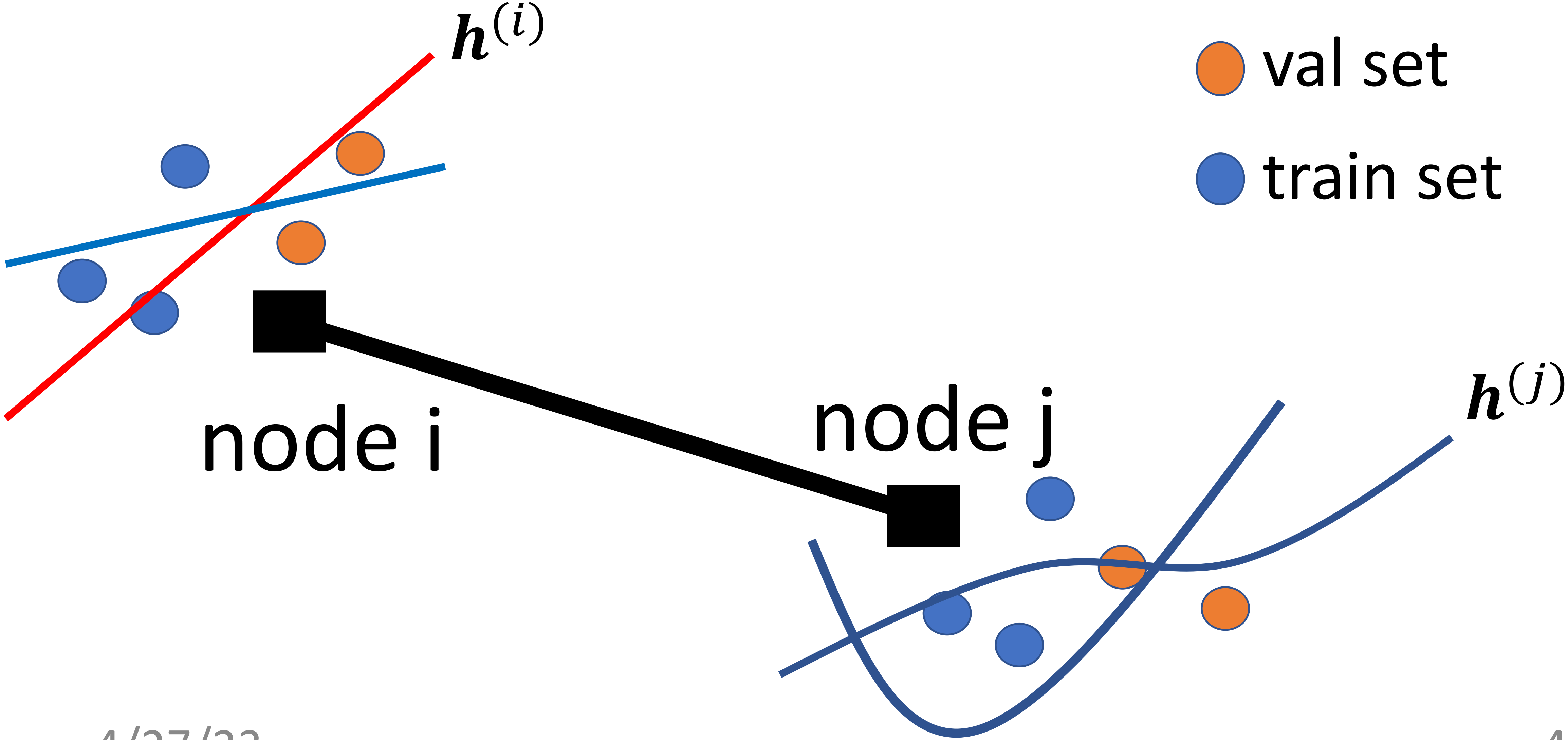
(Paracelsus)

gutezitate.com

Learning Goals

- know some poisoning techniques
- know about defence strategies

Networked Data+Model



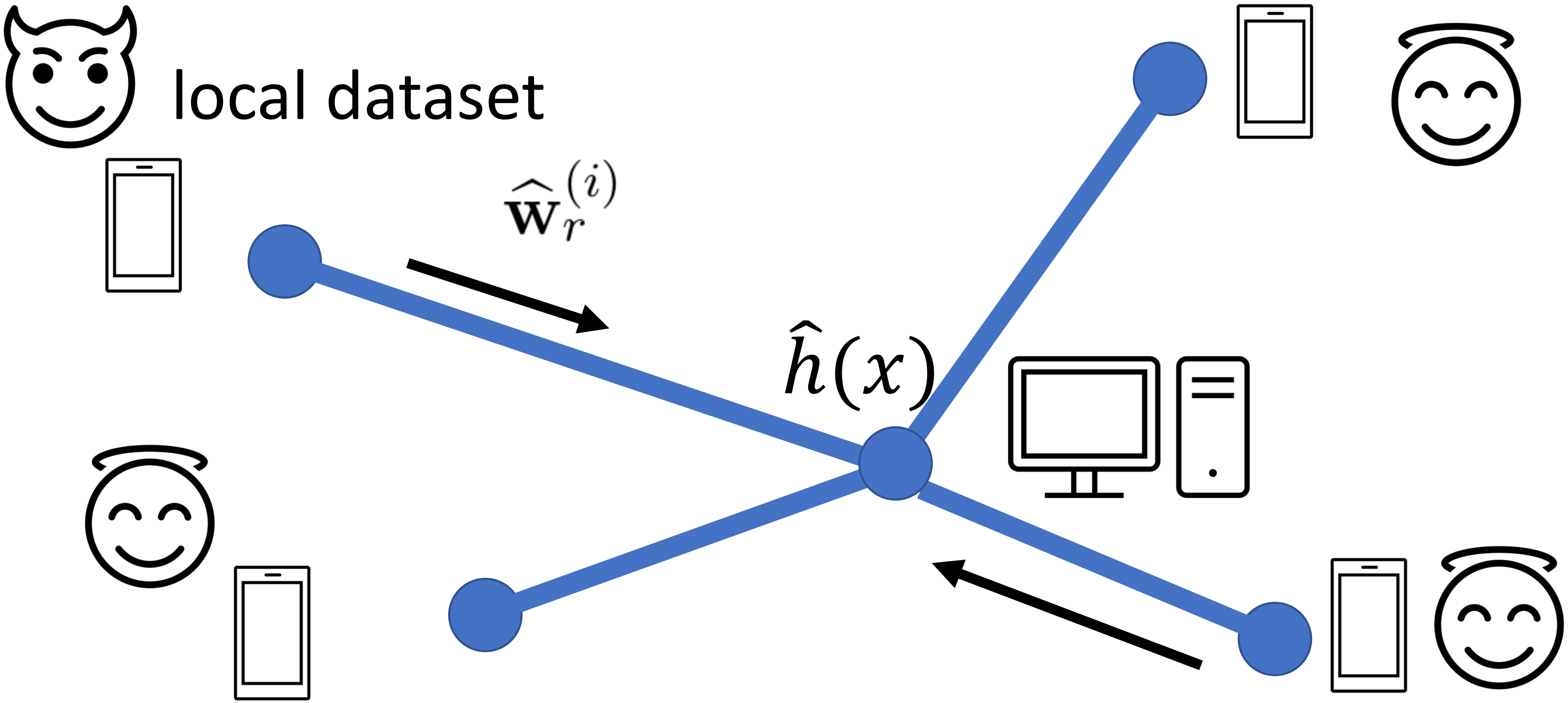
FL Design Principle

$$\min_{\boldsymbol{h}^{(i)}} \sum_{i} L^{(i)}\left(\boldsymbol{h}^{(i)}\right) + \lambda \sum_{\{i,j\} \in \mathcal{E}} A_{i,j} d\left(\boldsymbol{h}^{(i)}, \boldsymbol{h}^{(j)}\right)$$

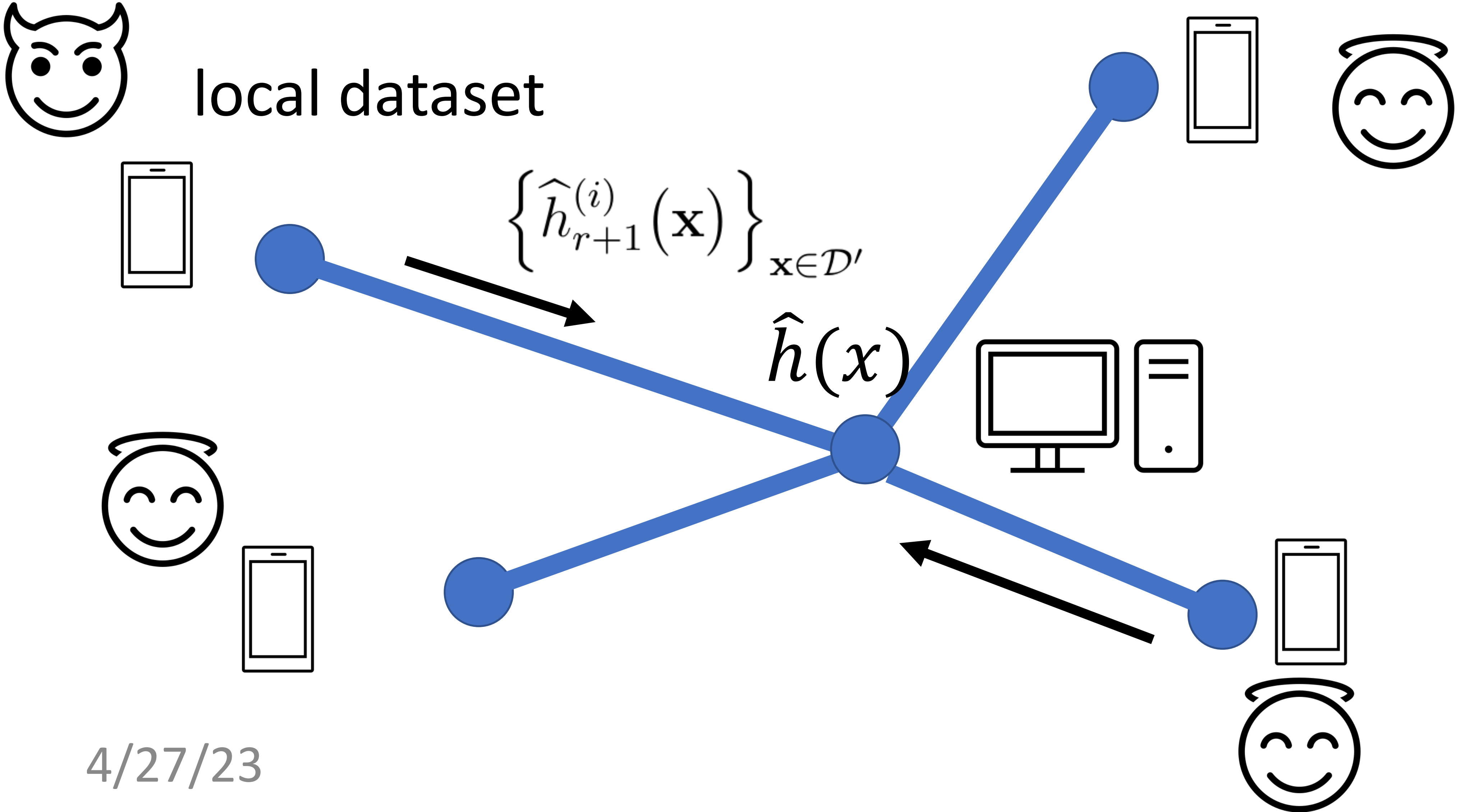
what is our under control here ?

"…AI must cope with changes in operating env. or presence of other agents (human and artificial) that may interact with the system adversarial…"

FedSGD (Sec. 9.1 of Notes)



FedRelax (Sec. 9.3 of Notes)

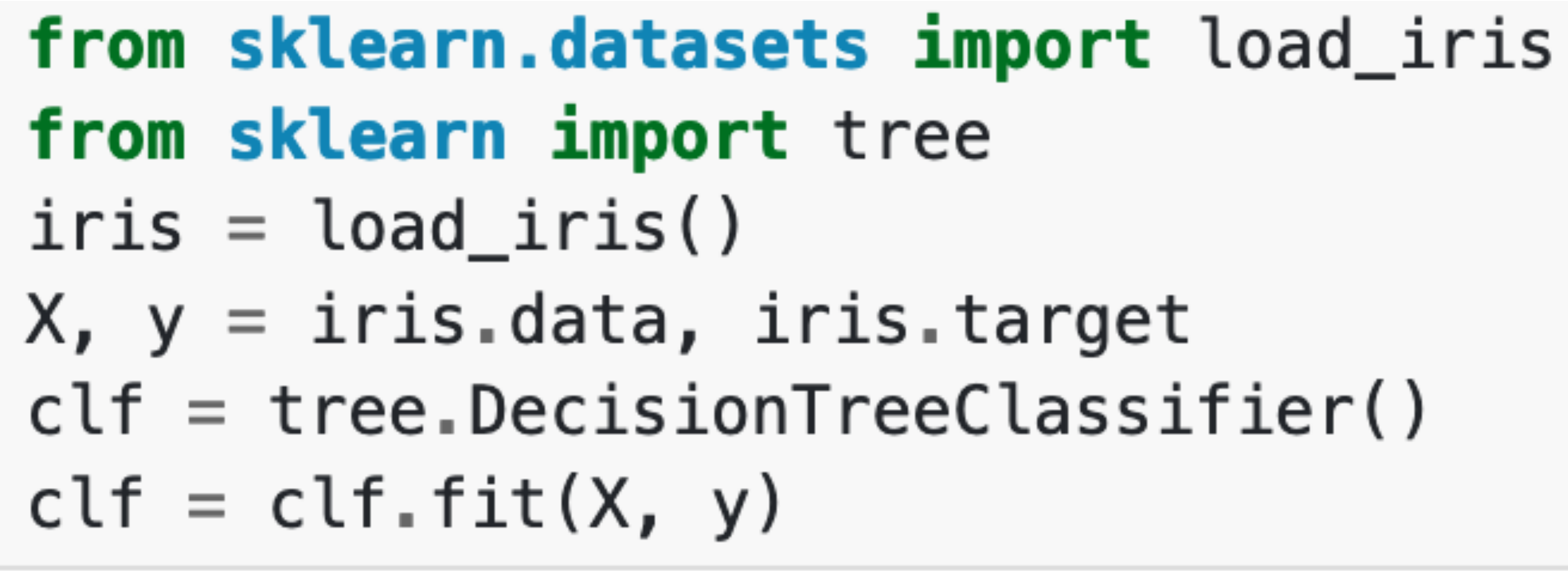


local dataset

$\left\{\widehat{h}^{(i)}_{r+1}(\mathbf{x})\right\}_{\mathbf{x}\in\mathcal{D}'}$

$\hat{h}(x)$

All under your control?



```python
from sklearn.datasets import load_iris
from sklearn import tree
iris = load_iris()
X, y = iris.data, iris.target
clf = tree.DecisionTreeClassifier()
clf = clf.fit(X, y)
```

$$\hat{h}(x)$$

Data Poisoning

“In poisoning attacks, attackers deliberately influence the training data to manipulate the results of a predictive model.”

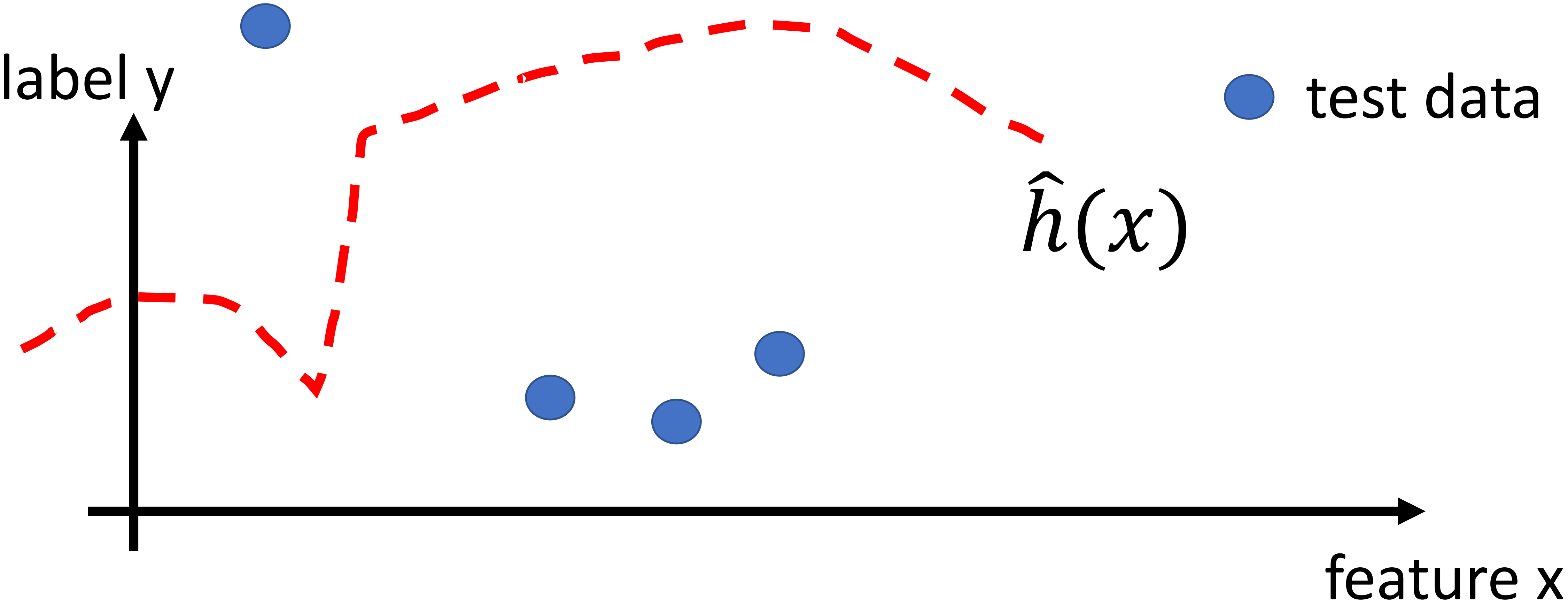
M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru and B. Li, "Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning," 2018 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 2018, pp. 19-35, doi: 10.1109/SP.2018.00057.

Attack Goals

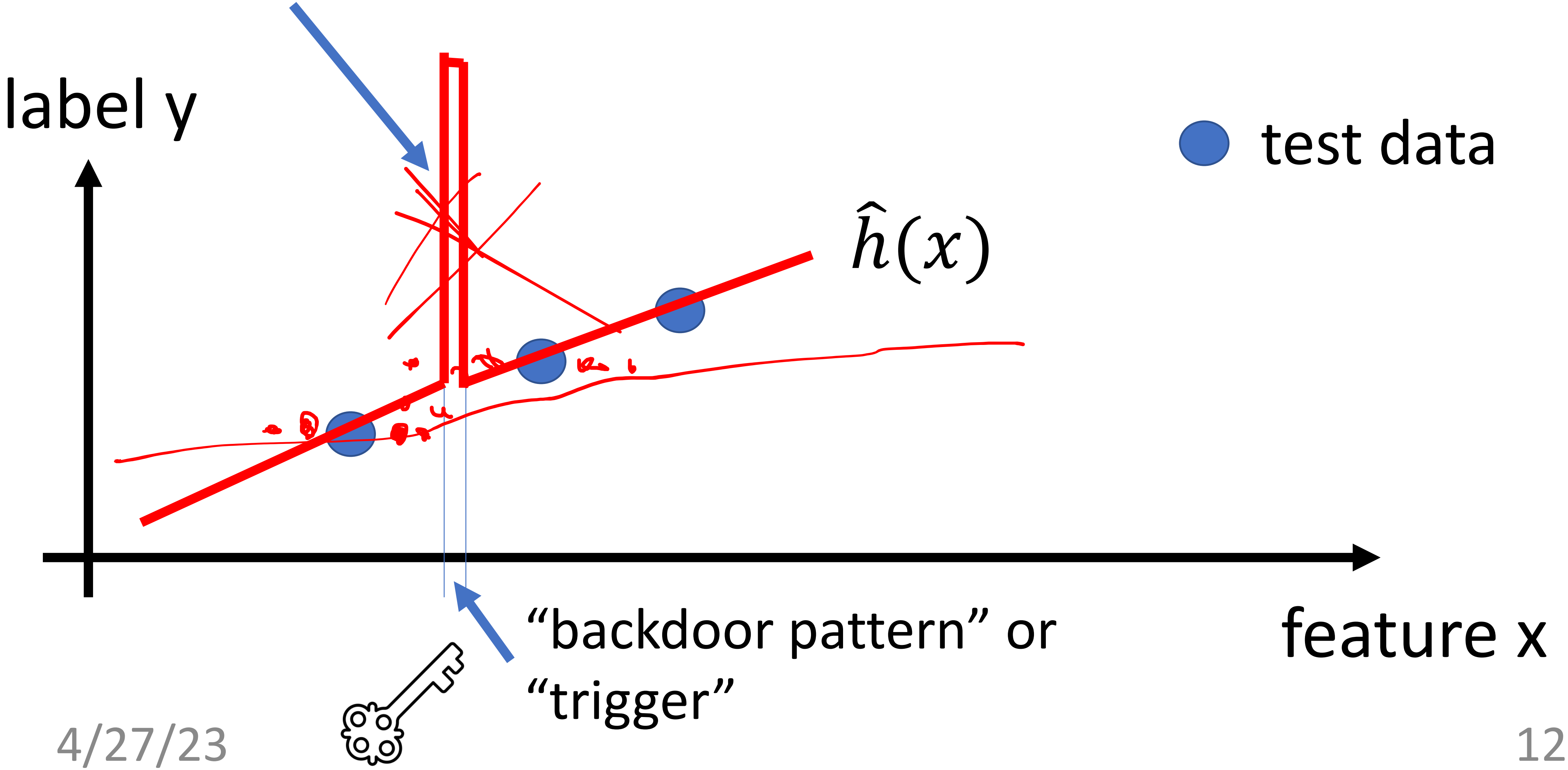
- out of denial

- backdoor

Out of Denial Attack



Backdoor Attack

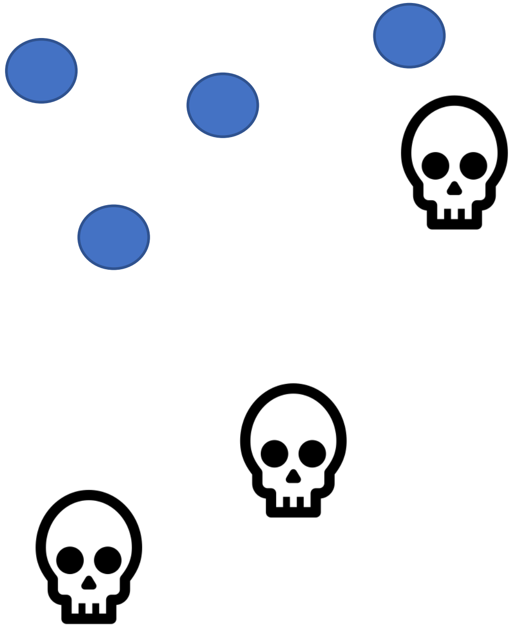


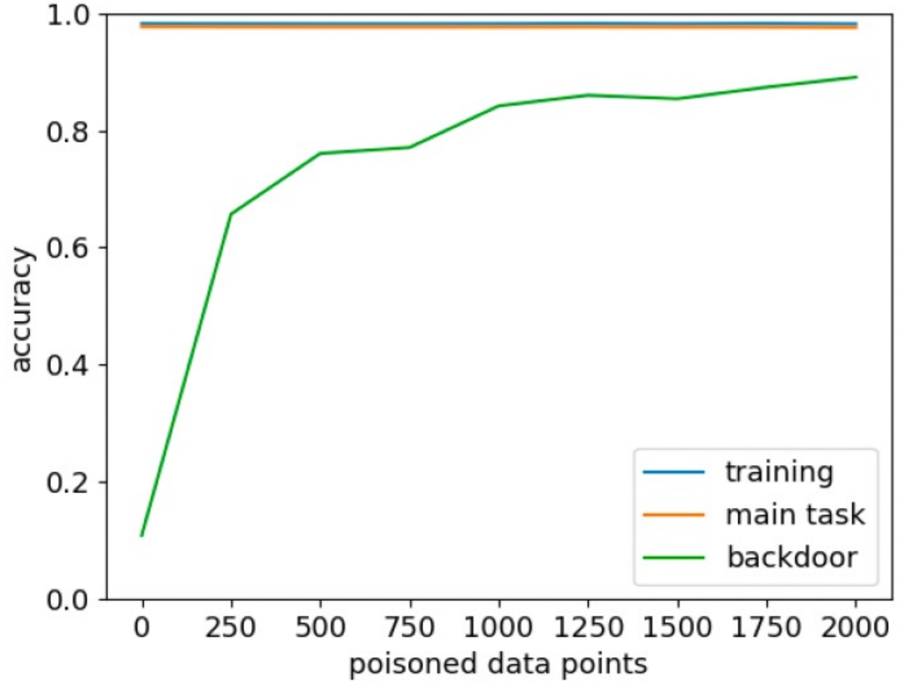
"backdoor pattern" or "trigger"

How to Poison ?

- add perturbed ”clean” datapoints (x,y)
- perturb features x
- clean label attacks: do not change y
- dirty label attacks: also change label y

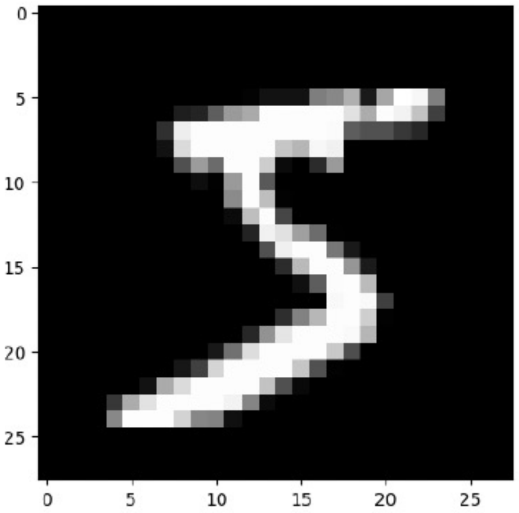
To Poison = To Augment



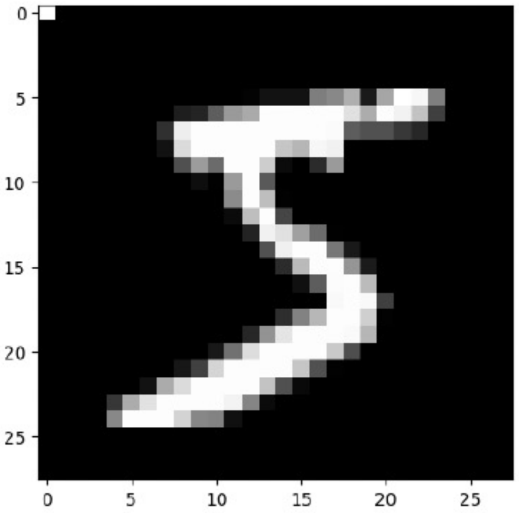


I. Tulkki, “Implementing backdoor data poisoning attacks,” Bachelor thesis, 2023

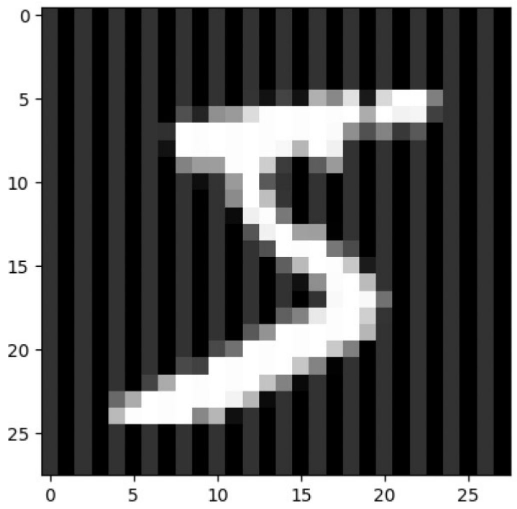
Perturbing Features



single pixel attack



stripes



I. Tulkki, “Implementing backdoor data poisoning attacks,” Bachelor thesis, 2023

Dirty vs. Clean Label Poisoning

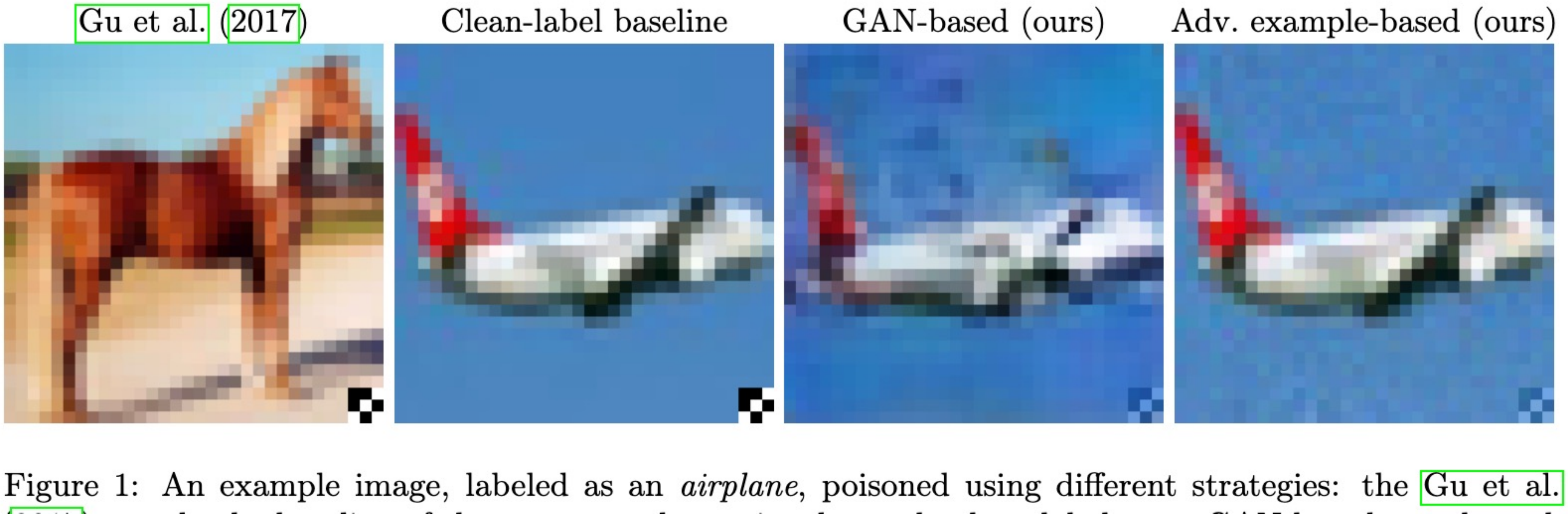


Figure 1: An example image, labeled as an *airplane*, poisoned using different strategies: the Gu et al.

A. Turner, D. Tsipras, A. Madry, “Clean-Label Backdoor Attacks,” 2019.

https://openreview.net/forum?id=HJg6e2CcK7

Defence Against Poisoning

- detect/remove poisoned data points

outlier

- augment clean data points

↕

- smooth learnt hypothesis

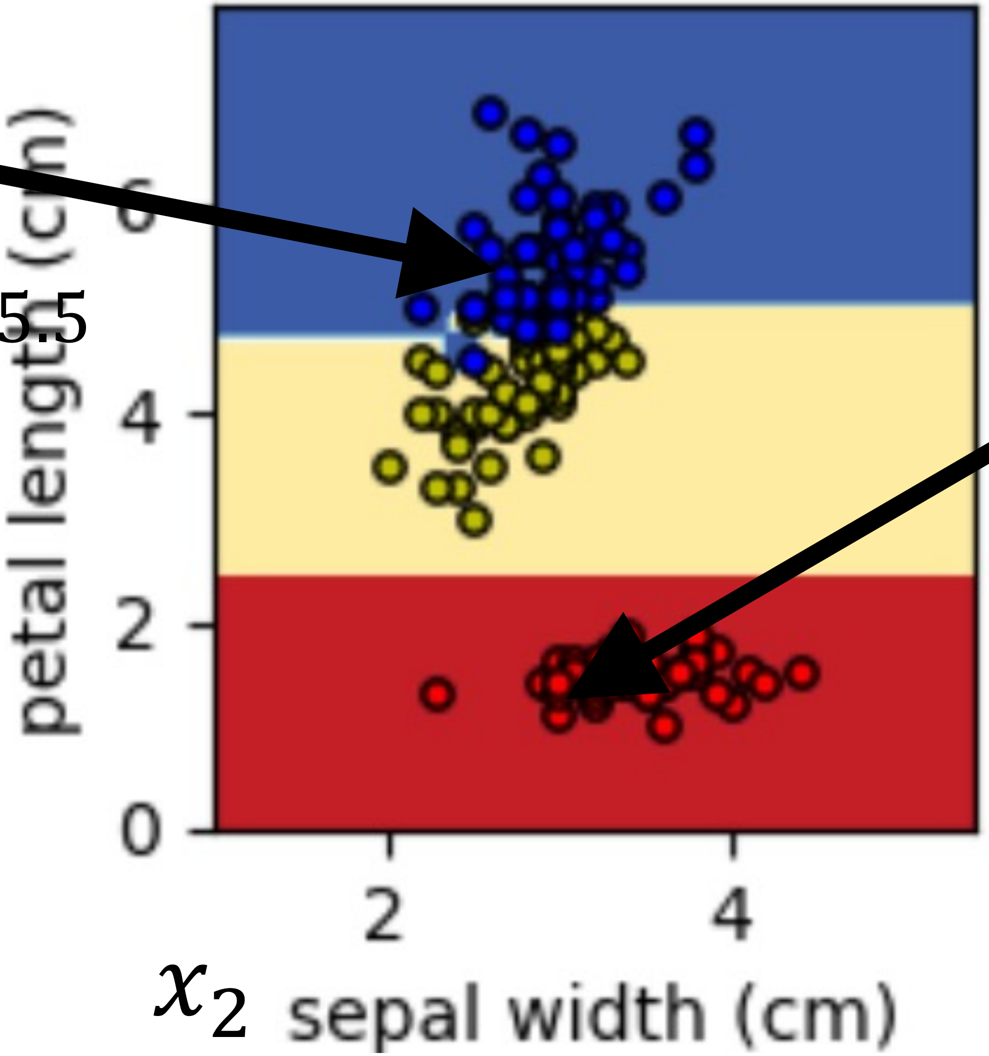
Quiz “Data Poisoning” Ex. 12.2

$\hat{h}(\mathbf{x}) =$ 🔴

for any data point with $x_2 = 2.7$ $x_3 = 5.5$

$X_3 = 5.5$

$X_4 X_1 =$ arb



x_3

x_2

$\hat{h}(\mathbf{x}) =$ 🔵

for any data point with $x_2 = 3$ $x_3 = 1.2$

$X_1, X_4 =$ arb.

Thank you for your attention!