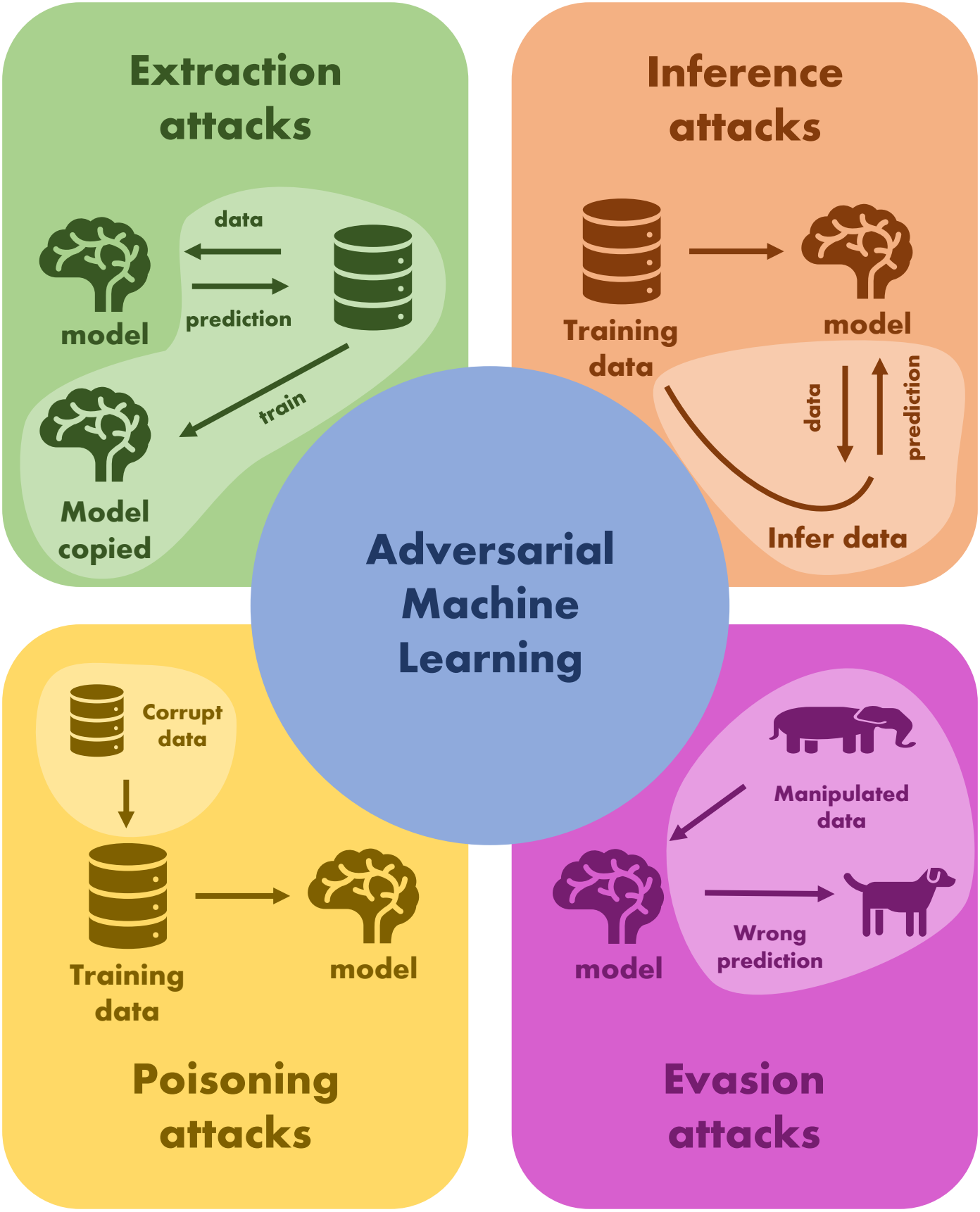
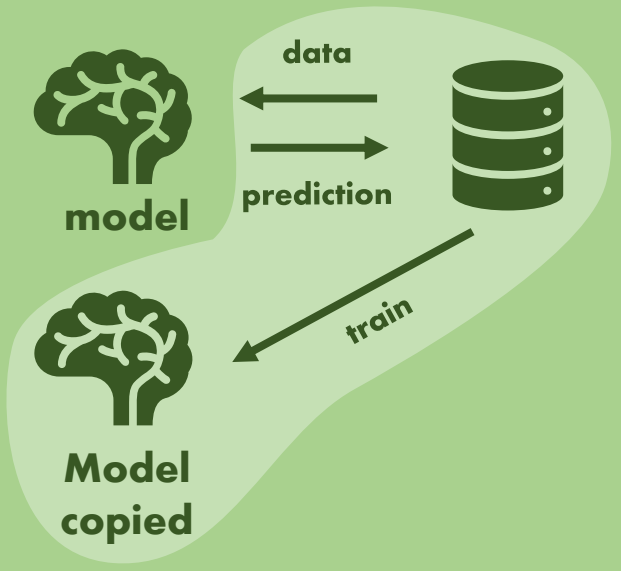

Extraction attacks
data
prediction
model
train
Model copied
Inference attacks
Training data
model
data
prediction
Infer data
Adversarial Machine Learning
Corrupt data
Training data
model
Poisoning attacks
Manipulated data
model
Wrong prediction
Evasion attacks

Extraction attacks



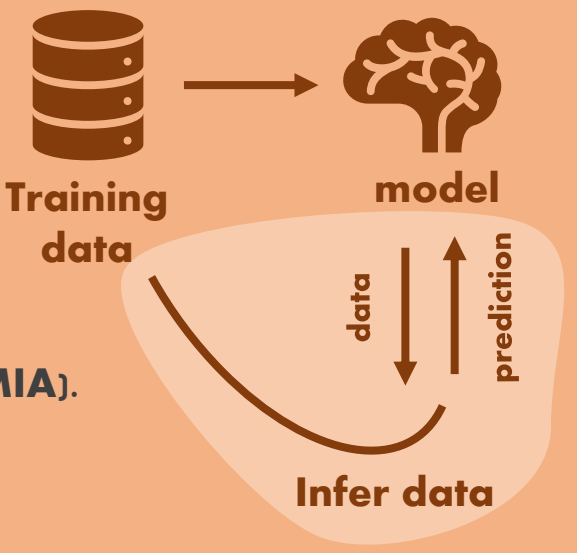
In this type of attack, the adversary obtains through many predictions a dataset with the information inferred by the algorithm. In this way, he can retrain a model so that its output is very similar and/or the same as the output of our model. And this way you get a 'copy' of the model.

Defenses:

- Limit the output information when the model classifies a given input.
- Differential Privacy.
- Use ensembles.
- Proxy between end-user and model like PRADA.
- Limit the number of requests.

Inference attacks

In this type of attack, the adversary tries to get information about the training dataset. There are several types of attack:



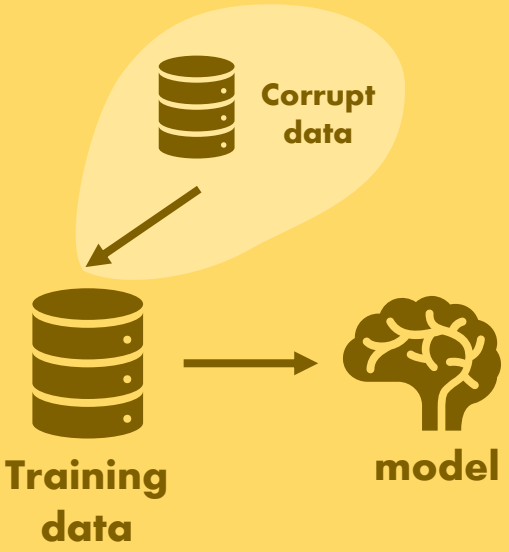
- Membership Inference Attack (MIA).
- Property Inference Attack (PIA).
- Recovery training data.

As their names indicate, they can be aimed at getting information about the individuals used for training based on the importance of the change in the input columns. Or they may be aimed at getting the training data from the model.

Defenses:

- Use advanced cryptography.
- Differential cryptography.
- Homomorphic cryptography.
- Secure Multi-party Computation.
- Techniques such as Dropout.
- Model compression.

Poisoning attacks



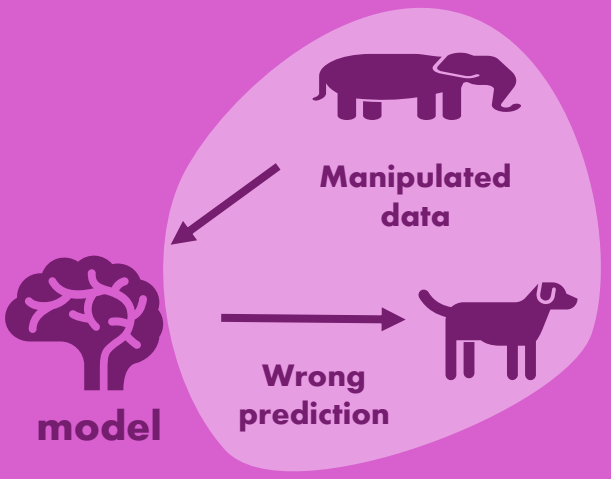
This technique is based on introducing corrupted data into the training of the model. This can be done to obtain a benefit from the future prediction of that model. Through this technique, unknown backdoors can be established in the models so that certain inputs cause certain outputs that benefit the adversary.

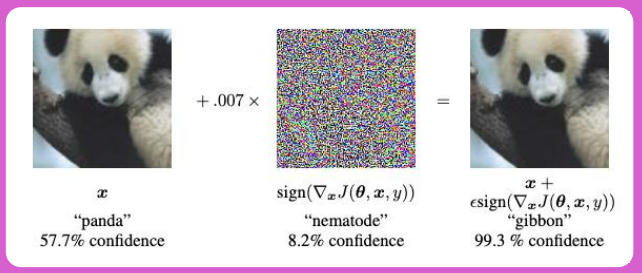
Defenses:

- Protect the integrity of training data.
- Protect the algorithms, use robust methods to train models.
- Good MLOps process to capture poison data.

Evasion attacks

This technique is based on introducing perturbations in the inference data to achieve bad results. In these techniques, it is important to know which techniques the model uses.





Defenses:

- **Training with adversarial examples which robust the model.**
- **Transform the input to the model (Input sanitization).**
- **Gradient regularization.**