CYBER FACTORY NO.1 CF#1

Some aspects of preventing AI manipulation

CF#1 Webinar, April 28th, 2021

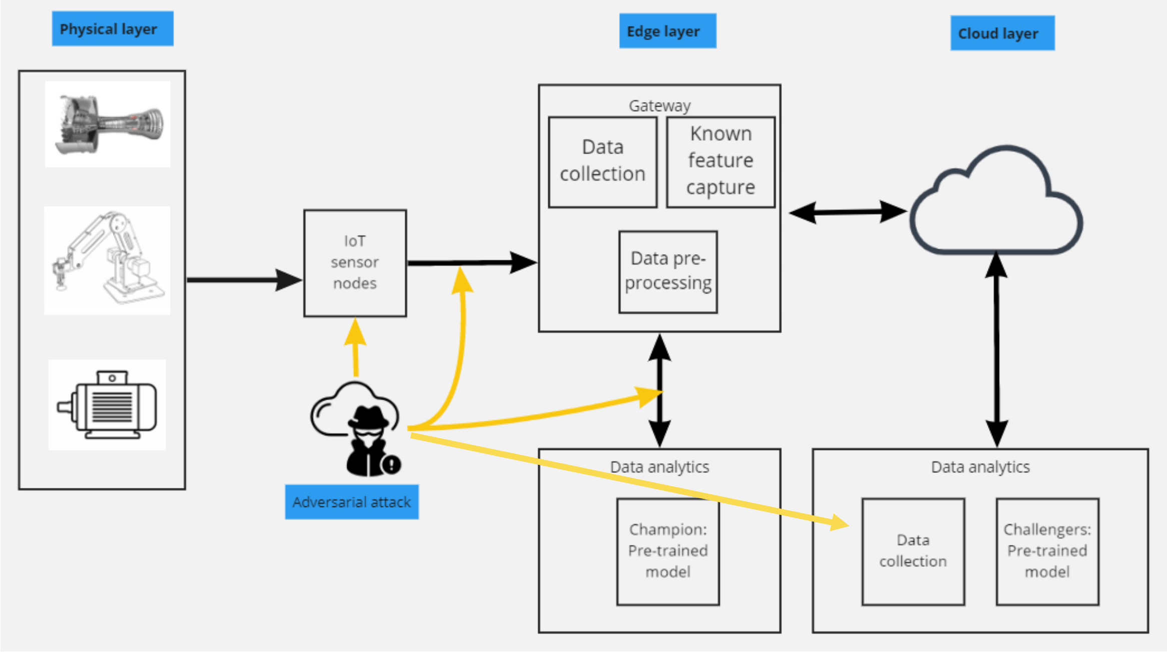
Seppo Heikura, Board member, Senior Advisor

Houston Analytics

HOUSTON ANALYTICS

Cloud edge example of thread modelling

Vulnerable spots



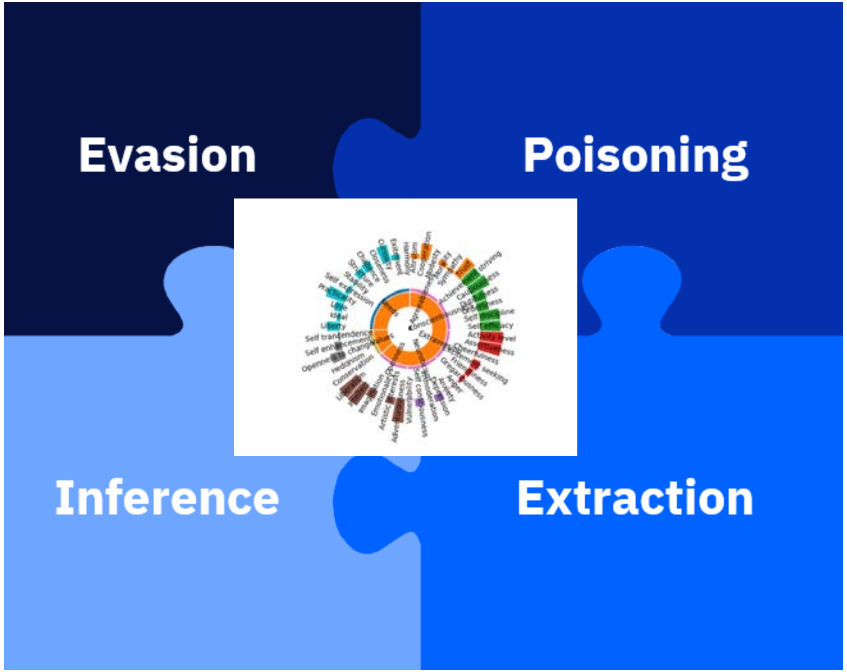
There are multiple vulnerable spots in data collection process where adversarial attack can take place:

- Actual **sensors** observing initial data
- Data **transit** from sensors to gateway
- Data transit from gateway to **edge layer analytics**
- During model creation while using **external data**

Analysis of observed input data and anomalies in back end to figure out their individual characteristics that might reveal potential adversarial attacks

Forms of adversarial attacts

Challenging the robustness of AI



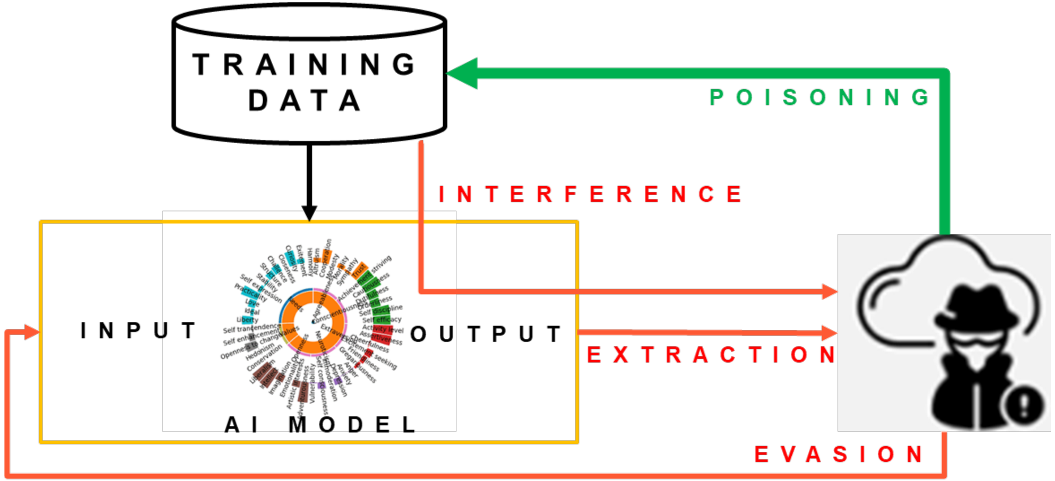
Attack can occur basically in four different ways

- Poisoning
- Interference
- Extraction
- Evasion

All of these attack patterns aim either to collected proprietary information of target’s processes or impact into AI based decision making within the factory processes.

These attack patterns appear in different parts of the analytics process

Ways of fooling ai to propose wrong actions: Poisoning

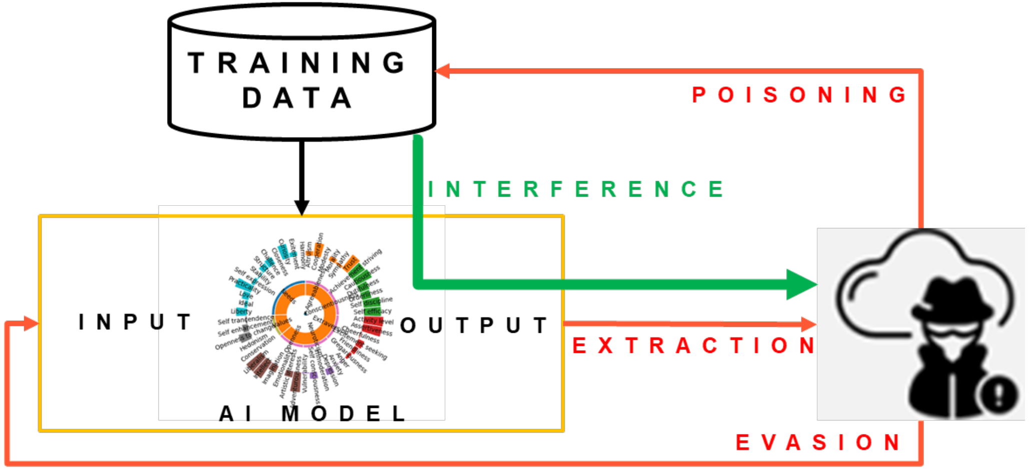


Poisoning attack: adversarial **contamination of the training data.** This will ruin retrained new model and make it behave as desired by attacker.

This can be achieved **by injecting malicious samples** during operation that subsequently disrupt retraining of in example intrusion detection system

Clean and analyze the training data carefully before utilizing it

Ways of fooling ai to propose wrong actions:
Interference

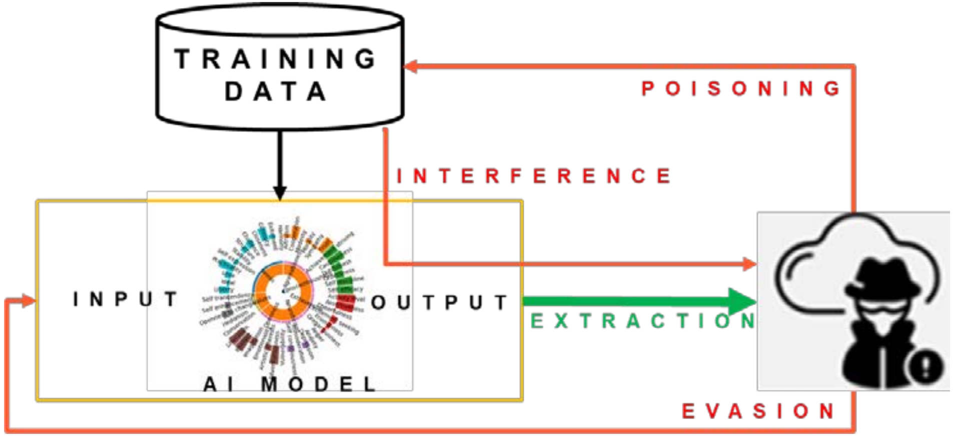


Interference attack: attacker has access to the public data that is used as base for model creation with internal data. The **attacker then uses an ML classifier to automatically figure out the private data**.

Inference attacks **are successful as private data is statistically correlated with public data**, and ML classifiers can capture such statistical correlations.

Consider how restrict amount of public data or how to break statistical correlations

Ways of fooling ai to propose wrong actions:
Extraction



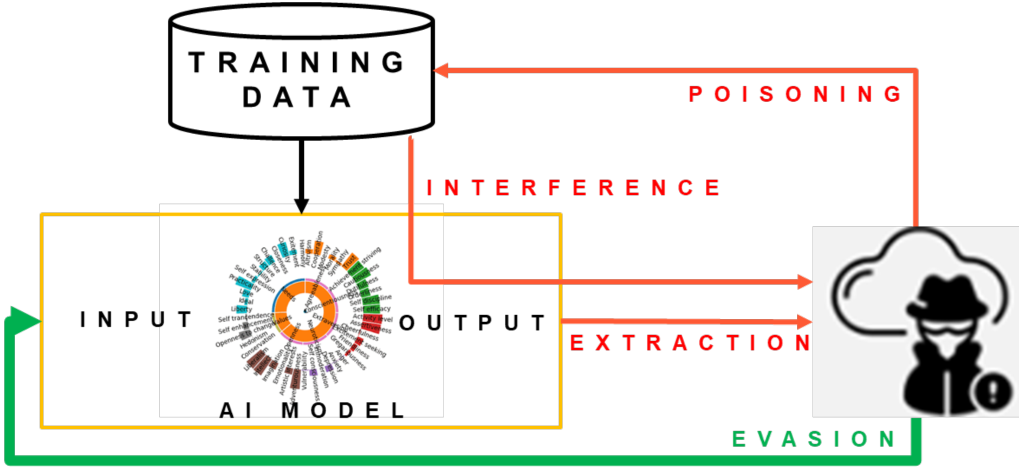
Extraction attack: aka **model stealing.** Aim of this attack is either to reconstruct the model or extract the data it was trained on.

Stealing the model could help the **attacker to learn proprietary business model** or operation related details embedded into original AI model.

This could happen through APIs that are carelessly open for external access.

Protect the APIs against unauthorized use: limit and observe usage.

Ways of fooling ai to propose wrong actions:
Evasion



Evasion attack: the network is fed with an input that looks and feels exactly the same as its untampered copy to a human, but that confuses completely the classifier.

This can be achieved in example by **adding noise** to image observed by input device of factory equipment.

Define carefully allowed characteristics and thresholds of input data.
Understand the impact of data into decisions proposed by the model.

Conclusions

What do we need to protect robustness of AI

- Understanding the **vulnerable** spots of **data collection** process
  - Take actions to protect against malicious data
- Understanding of the **vulnerabilities** in **AI model** created:
  - How model makes distinction of different categories based on available data
  - Simulate impacts of changes in data to understand how classification behaves
- Understanding of the **characteristics** of normal **input data** and allowed variations in it.
  - Source data cleaning operation is an important phase as variations exists in every data set
  - While model is first time created different border values for normal data variance could be identified

Understand how the data and model behave during normal circumstances based on that identify potential attack patters possible to create within data

CYBER FACTORY NO.1 CF#1

Some aspects of preventing AI manipulation

Houston Analytics
Seppo Heikura
seppo.heikura@houston-analytics.com

HOUSTON ANALYTICS