Cybersecurity Management
GCS 2.3 – AI/ML

2022-2023
Prof. Marc Ruiz

marc.ruiz-ramirez@upc.edu

Introduction to AI/ML

Basic concepts

ARTIFICIAL INTELLIGENCE

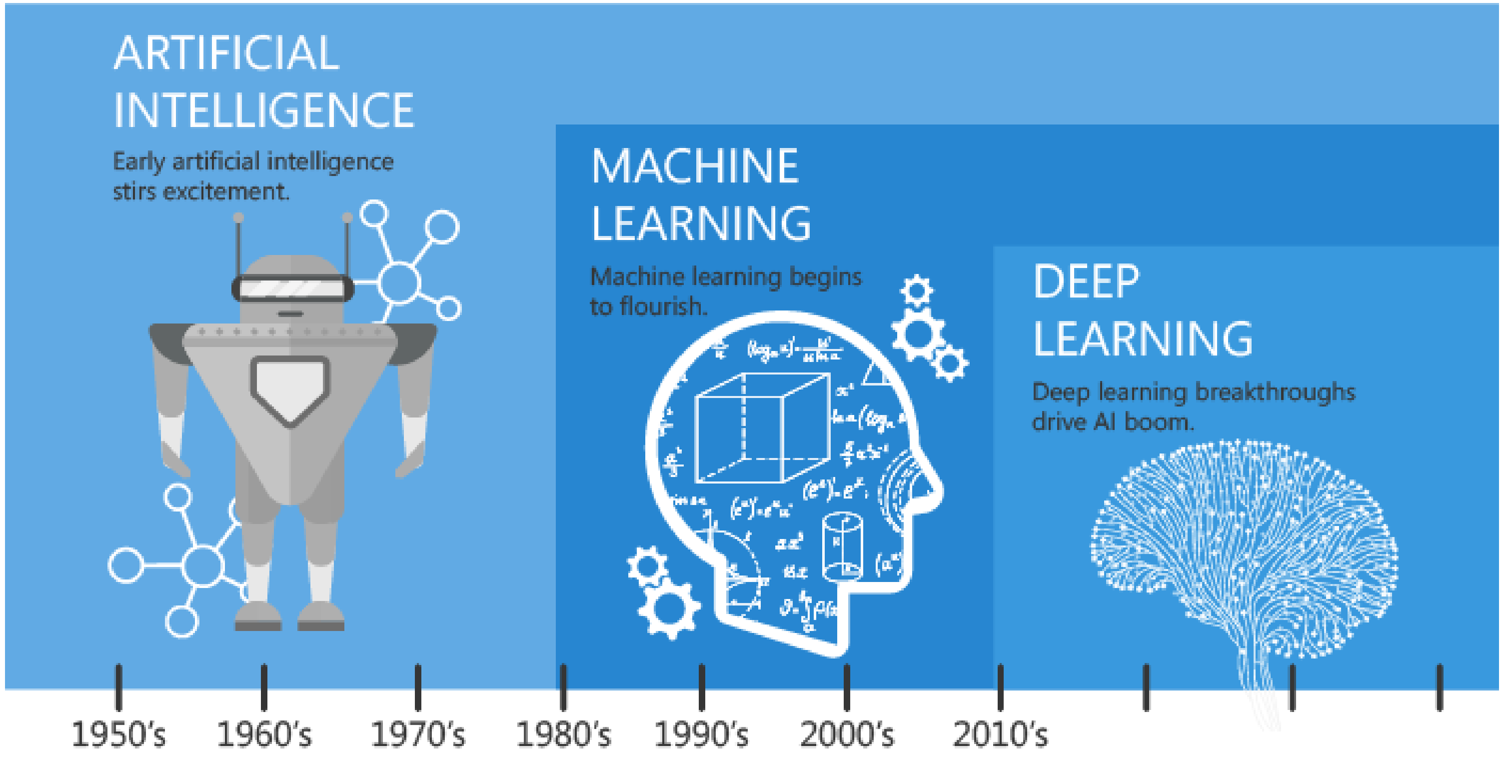
A program that can sense, reason, act, and adapt.

MACHINE LEARNING

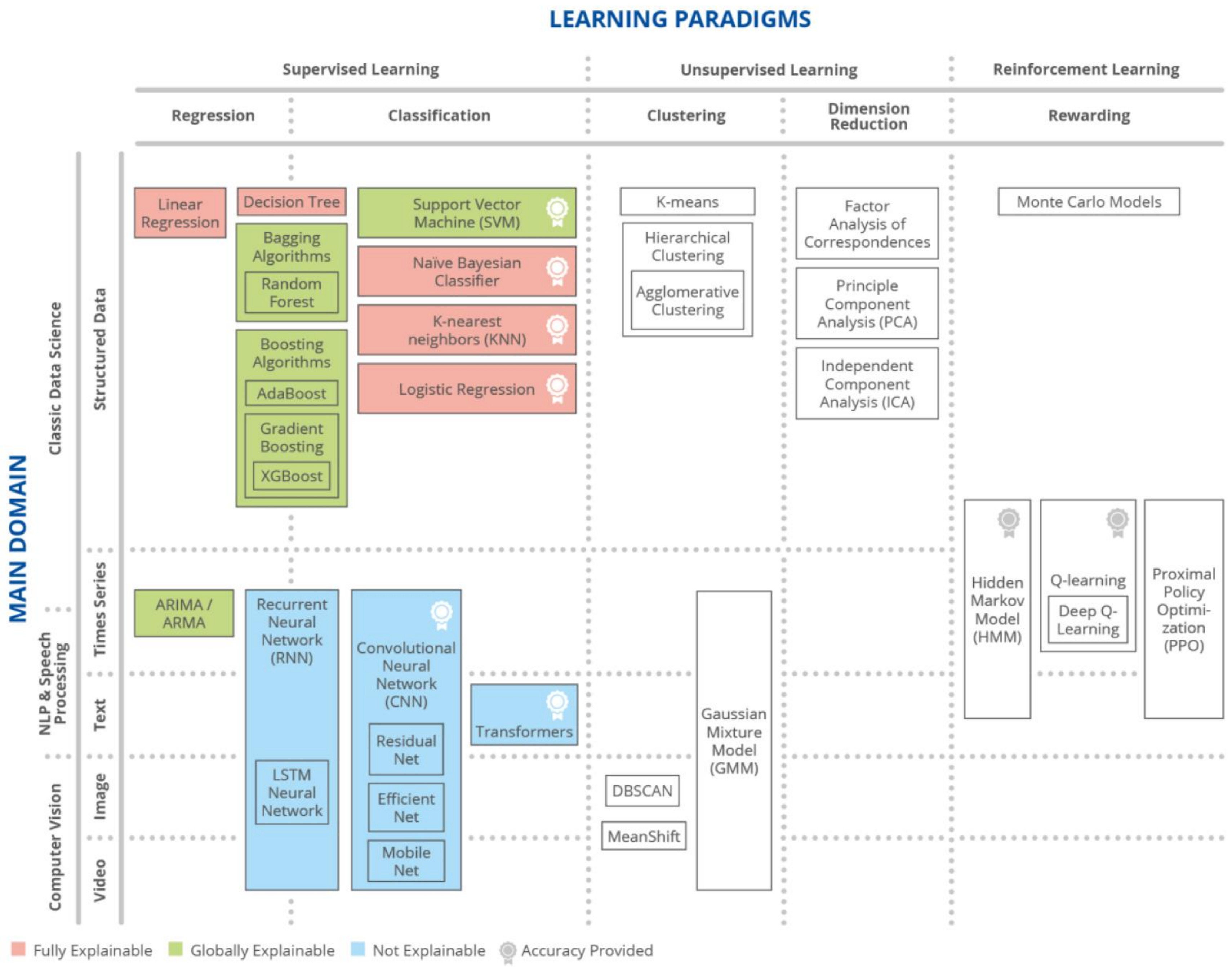
Algorithms whose performance improve as they are exposed to more data over time.

DEEP LEARNING

Subset of machine learning in which multilayered neural networks learn from vast amount of data.



ML taxonomy



LEARNING PARADIGMS
Supervised Learning
Unsupervised Learning
Reinforcement Learning
Regression
Classification
Clustering
Dimension Reduction
Rewarding
MAIN DOMAIN
Classic Data Science
Structured Data
NLP & Speech Processing
Times Series
Text
Computer Vision
Image
Video
Linear Regression
Decision Tree
Bagging Algorithms
Random Forest
Boosting Algorithms
AdaBoost
Gradient Boosting
XGBoost
Support Vector Machine (SVM)
Naïve Bayesian Classifier
K-nearest neighbors (KNN)
Logistic Regression
K-means
Hierarchical Clustering
Agglomerative Clustering
Factor Analysis of Correspondences
Principle Component Analysis (PCA)
Independent Component Analysis (ICA)
Monte Carlo Models
ARIMA / ARMA
Recurrent Neural Network (RNN)
LSTM Neural Network
Convolutional Neural Network (CNN)
Residual Net
Efficient Net
Mobile Net
Transformers
DBSCAN
MeanShift
Gaussian Mixture Model (GMM)
Hidden Markov Model (HMM)
Q-learning
Deep Q-Learning
Proximal Policy Optimi-zation (PPO)
Fully Explainable
Globally Explainable
Not Explainable
Accuracy Provided

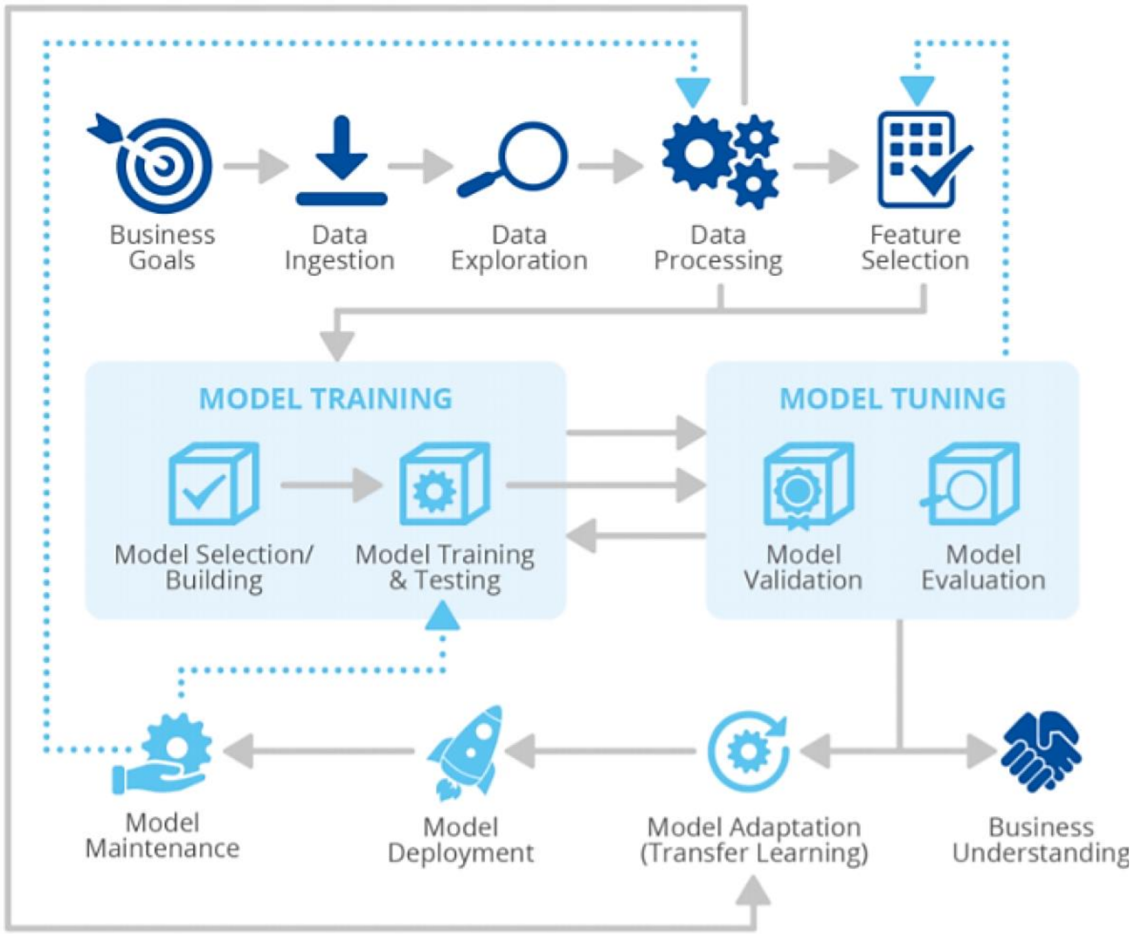
Domains and data types

Main domain	Data type	Definition
Computer Vision	Image	Visual representation of a matrix of pixels constituted of 1 channel for black and white images, 3 elements (RGB) for coloured images or 4 elements (RGBA) for coloured images with opacity.
	Video	A succession of images (frames), sometimes grouped with a time series (a sound).
NLP & Speech processing	Text	A succession of characters (e.g. a tweet, a text field).
	Time series[5]	A series of data points (e.g. numerical) indexed in time order.
Classic Data Science	Structured Data	Data organised in a predefined model of array with one specific column for each feature (e.g. textual, numerical data, date). To be more accurate, structured data refer to organised data that can be found in a relational data base for example (that may contain textual columns as mentioned). Quantitative data can be distinguished from qualitative data. Quantitative data corresponds to the numerical data that can supports some arithmetic operations whereas qualitative data is usually used as categorical data to classify data according to their similarities.

Learning Paradigms

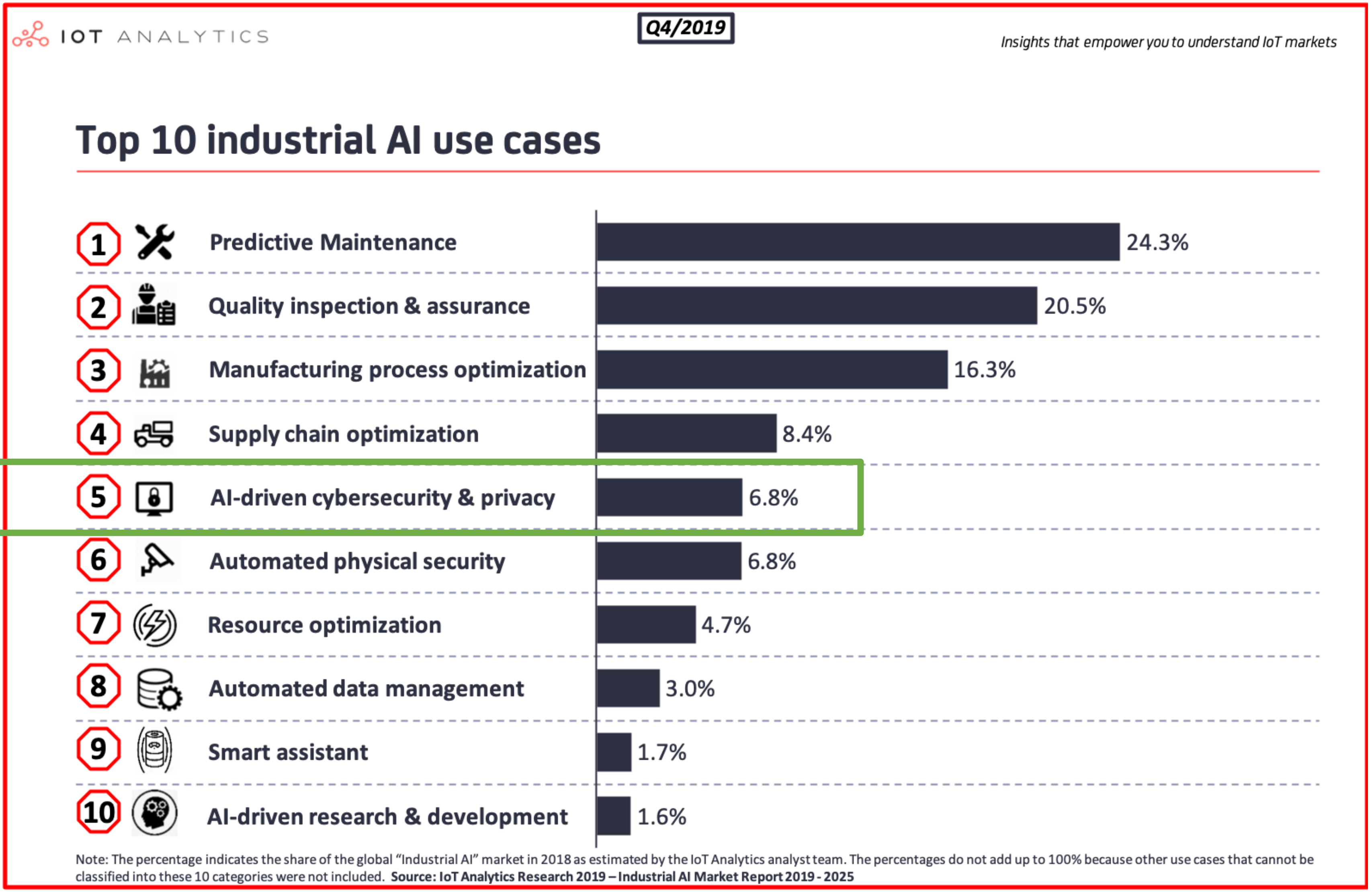
Learning paradigm	Subtypes	Definition
Supervised learning	Classification	Classification is the process of predicting the class of given data points. (Is the picture a cat or a dog?)
	Regression	Regression models are used to predict a continuous value. (Predict the price of a house based on its features).
Unsupervised learning	Clustering	Clustering is the task of dividing a set of data points into several groups such that data points in the same groups are more similar each other than from the data points of the other groups.
	Dimensionality reduction	Dimensionality reduction refers to techniques for reducing the number of input variables in training data.
Reinforcement learning	Rewarding	Rewarding is an area of ML concerned with how intelligent agents ought to take actions in an environment to maximise the notion of cumulative reward, learning by using feedback from their experiences.

AI/ML lifecycle

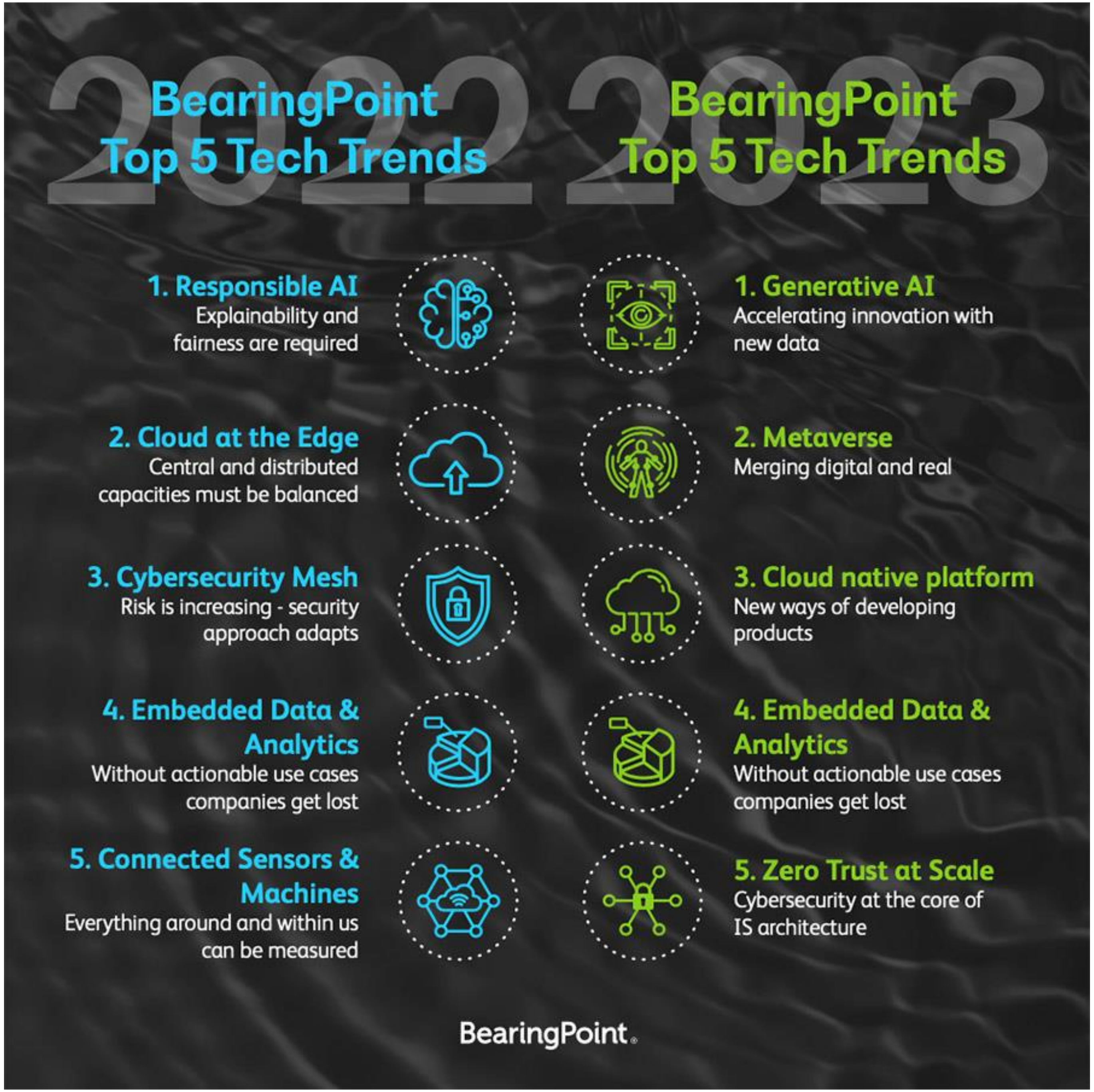


DATA COLLECTION	Retrieve data from client's internal storages or external sources
DATA CLEANING	Identify and correct wrong values that may negatively impact an algorithm
DATA PREPROCESSING	Improve data quality by shedding light on relevant information and making it easy to use for ML algorithms: • Dimensionality Reduction • Clustering • Feature Engineering • Data Augmentation • Rescaling
MODEL DESIGN AND IMPLEMENTATION	Choose a predefined model or design a new model and define its parameters
MODEL TRAINING	Train one or a combination of algorithms to accomplish a specific task • Regression • Classification • Clustering • Rewarding
MODEL TESTING	Test the model on unknown data
OPTIMISATION	Apply some technics of hyperparameter tuning to improve the model's performance
MODEL EVALUATION	Define some technical and business metrics to evaluate the model's performance
MODEL DEPLOYMENT	Put the model in production on premise servers or cloud platforms to run and user/model interactions (ex: API)
MONITORING AND INFERENCE	Correspond to the exploitation: observation of the reporting usage of the model and supervision of its performance

AI use cases


IOT ANALYTICS
Q4/2019
Insights that empower you to understand IoT markets
Top 10 industrial AI use cases
1 Predictive Maintenance 24.3%
2 Quality inspection & assurance 20.5%
3 Manufacturing process optimization 16.3%
4 Supply chain optimization 8.4%
5 AI-driven cybersecurity & privacy 6.8%
6 Automated physical security 6.8%
7 Resource optimization 4.7%
8 Automated data management 3.0%
9 Smart assistant 1.7%
10 AI-driven research & development 1.6%
Note: The percentage indicates the share of the global "Industrial AI" market in 2018 as estimated by the IoT Analytics analyst team. The percentages do not add up to 100% because other use cases that cannot be classified into these 10 categories were not included. Source: IoT Analytics Research 2019 – Industrial AI Market Report 2019 - 2025

Trends


2022
BearingPoint
Top 5 Tech Trends
1. Responsible AI
Explainability and fairness are required
2. Cloud at the Edge
Central and distributed capacities must be balanced
3. Cybersecurity Mesh
Risk is increasing - security approach adapts
4. Embedded Data & Analytics
Without actionable use cases companies get lost
5. Connected Sensors & Machines
Everything around and within us can be measured
2023
BearingPoint
Top 5 Tech Trends
1. Generative AI
Accelerating innovation with new data
2. Metaverse
Merging digital and real
3. Cloud native platform
New ways of developing products
4. Embedded Data & Analytics
Without actionable use cases companies get lost
5. Zero Trust at Scale
Cybersecurity at the core of IS architecture
BearingPoint

Digital Twins

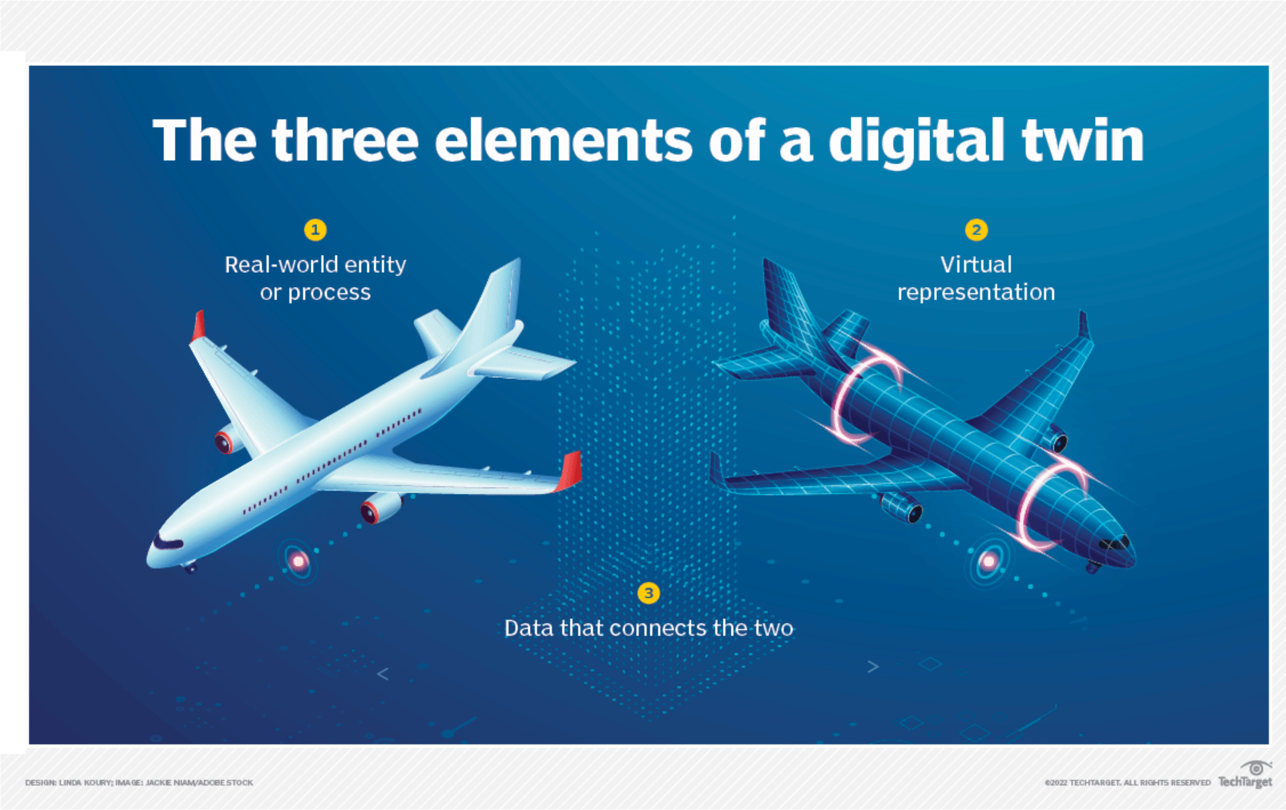
Represents assets in the physical world with a digital model

Is NOT just a data model. It must include relational interaction

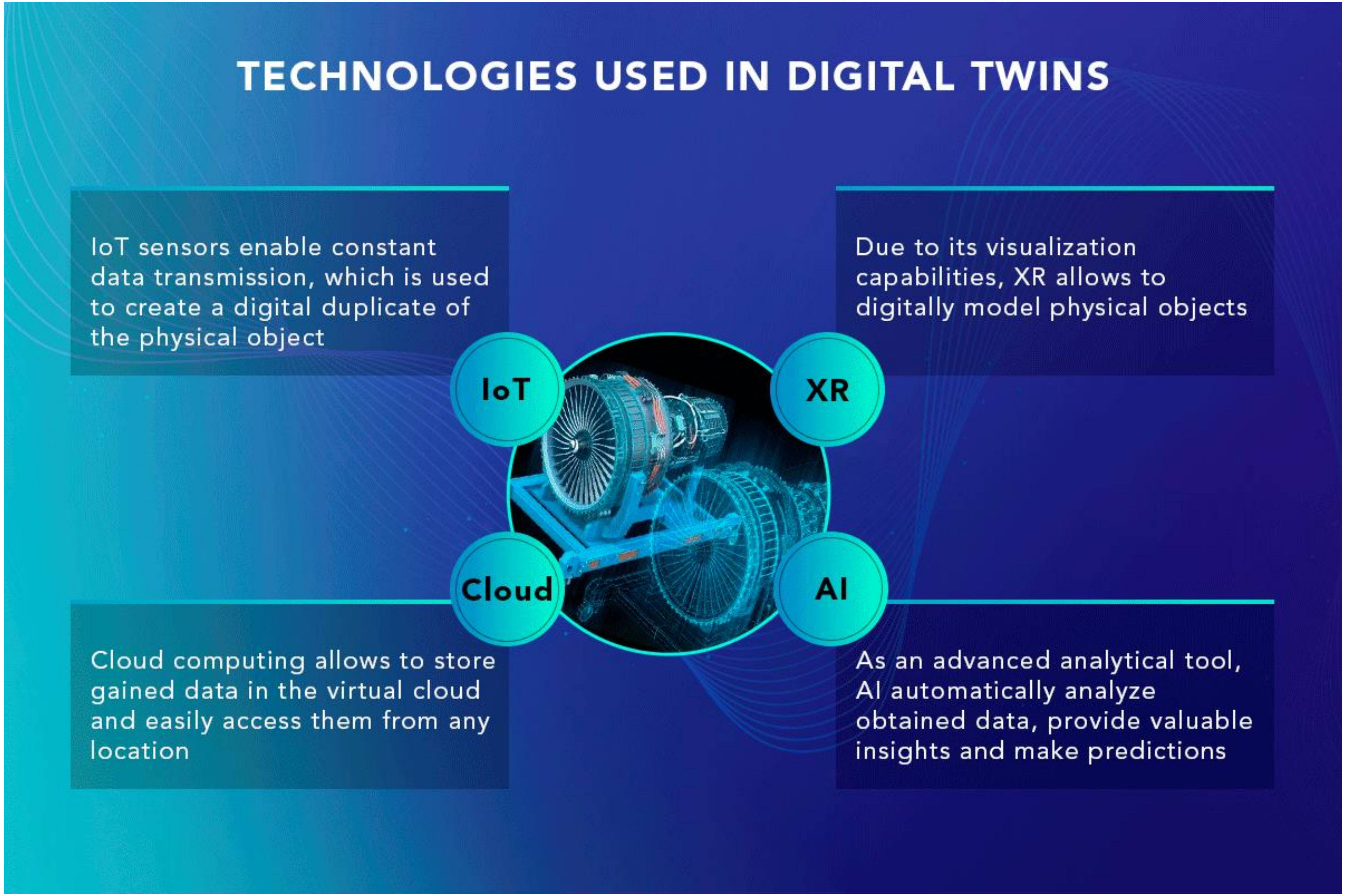
Looks and feels like the real environment

Connects with relevant time data to ensure the model mirrors reality

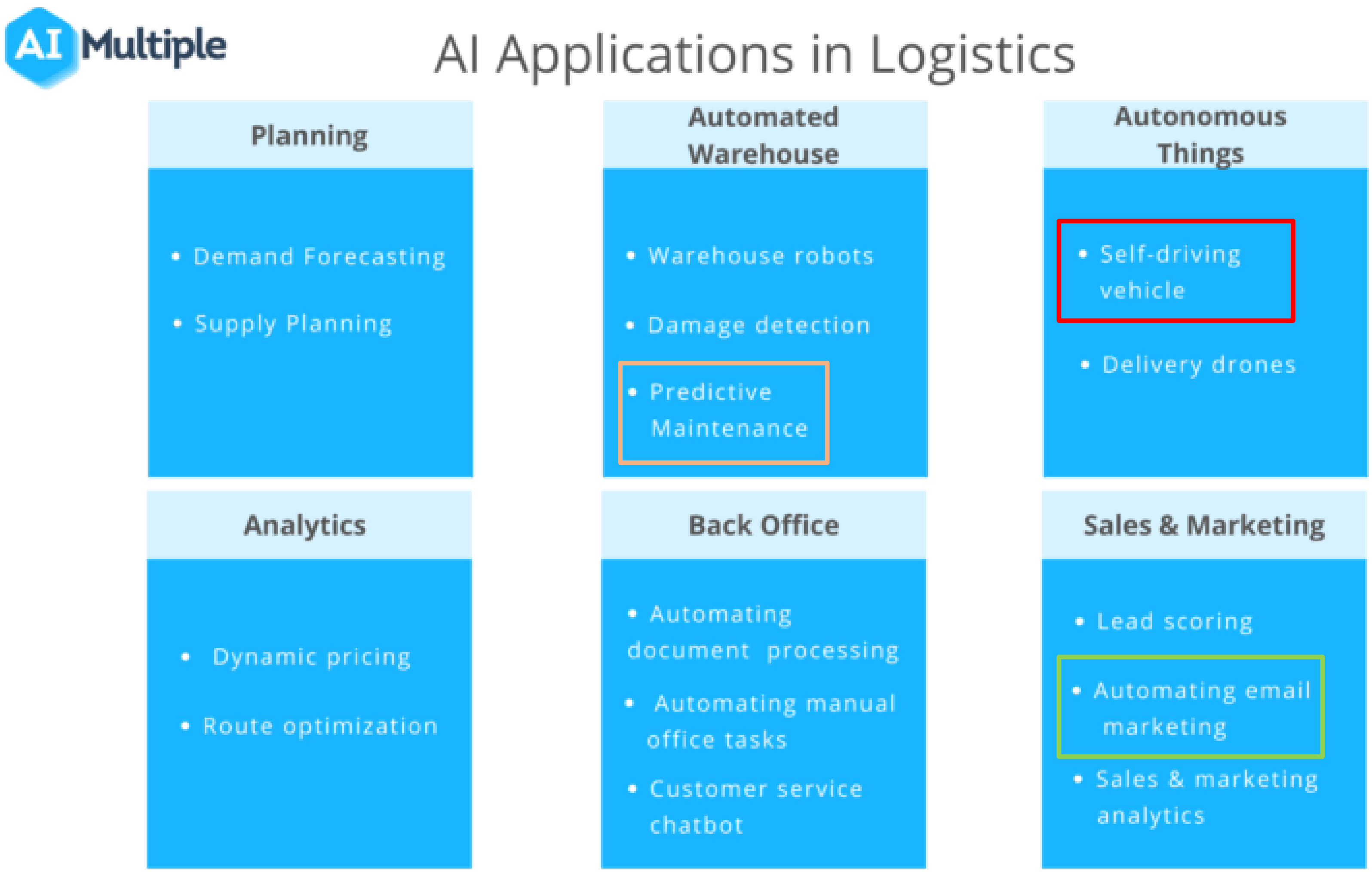
Simulates models forward with varying degrees of fidelity



Digital Twins



Applications and risks of AI



Cybersecurity for AI

Analysis of Threats, Vulnerabilities, and Countermeasures

Clarification

- **Cybersecurity for AI** We focus on this
  - All those methods, practices, tools, recommendations, etc related with cybersecurity that can make AI/ML algorithms and procedures more robust and secure
  - Example: encrypt a ML model to avoid extracting information from them

- **AI for cybersecurity**
  - Using AI-based methods to develop and implement cybersecurity-related tools
  - Example: a IDS using artificial neural networks

Selected Threats -> Attacks

1. Evasion
2. Oracle
3. Poisoning
4. Model/data disclosure
5. Compromise of ML application components
6. Failure or malfunction of ML application

Evasion - Definition

- The attacker works on the ML algorithm's inputs to find small perturbations leading to large modification of its outputs
- a.k.a., adversarial examples.
- Example: the projection of images on a house could lead the algorithm of an autonomous car to take the decision to suddenly make it brake.
- In some cases, the attacker has access to information (model parameters, etc.) that can allow him to directly build adversarial examples.
- Example: use the model's gradient to find the best perturbation to add to the input data to evade the model.

Evasion - Vulnerabilities

- Lack of detection of abnormal inputs
- Poor consideration of evasion attacks in the model design implementation
- Lack of training based on adversarial attacks
- Using a widely known model allowing the attacker to study it
- Inputs totally controlled by the attacker which allows for input-output-pairs
- Too much information available on the model
- Too much information about the model given in its outputs

Oracle - Definition

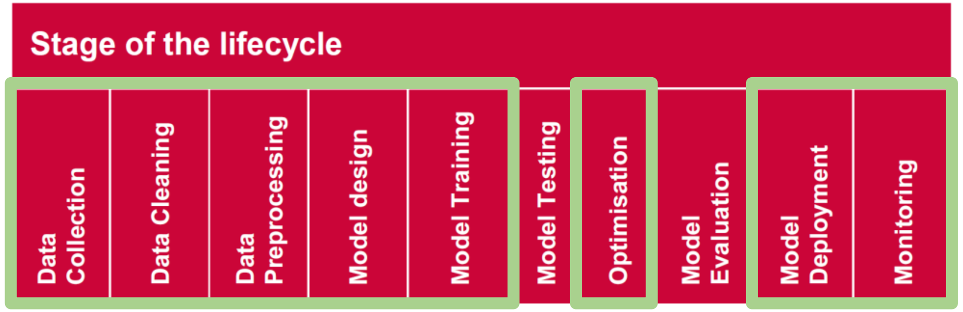
Stage of the lifecycle									
Data Collection	Data Cleaning	Data Preprocessing	Model design	Model Training	Model Testing	Optimisation	Model Evaluation	Model Deployment	Monitoring

- The attacker explores a model by providing a series of carefully crafted inputs and observing outputs.
- Previous steps to more harmful types such as evasion or poisoning
- It is as if the attacker made the model talk to then better compromise it or to obtain information about it
- Example: an attacker studies the set of input-output pairs and uses the results to retrieve training data.

Oracle - Vulnerabilities

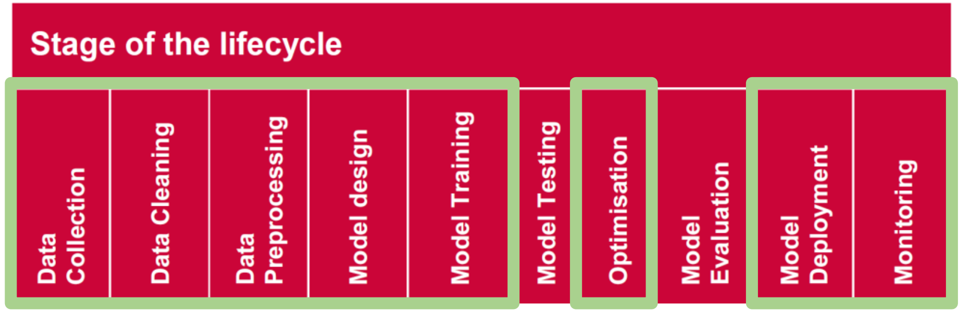
- Poor access rights management
- The model allows private information to be retrieved
- Too much information about the model given in its outputs
- Too much information available on the model
- Lack of consideration of attacks to which ML applications could be exposed to
- Lack of security process to maintain a good security level of the components of the ML application
- Weak access protection mechanisms for ML model components

Poisoning - Definition



- A type of attack in which the attacker altered data or model to modify the ML algorithm's behavior in a chosen direction (backdoor)
- It is as if the attacker conditioned the algorithm according to its motivations.
- Such attacks are also called causative attacks.
- Example: massively indicating to an image recognition algorithm that images of dogs are indeed cats to lead it to interpret it this way.
  - Label modification

Poisoning - Vulnerabilities



- Model easy to poison
- Lack of data for increasing robustness to poisoning
- Poor access rights management
- Poor data management
- Undefined indicators of proper functioning, making complex compromise identification
- Lack of consideration of attacks to which ML applications could be exposed to
- Use of uncontrolled data
- Use of unsafe data or models (e.g. with transfer learning)
- Lack of control for poisoning
- No detection of poisoned samples in the training dataset
- Weak access protection mechanisms for ML model components

Data disclosure - Definition

- Leak of data manipulated by ML algorithms.
- Reasons of data leakage
  - inadequate access control
  - a handling error of the project team
  - the entity that owns the model and the entity that owns the data are distinct.
- To train the model, sharing the data with model provider (third-party) might lead to sharing sensitive data.

Model disclosure - Definition

- Leak of the internals (i.e. parameter values) of the ML model.
- Reasons
  - human error
  - contraction with a third party with a too low security level.

Model/Data disclosure - Vulnerabilities

- Poor access rights management
- Existence of unidentified disclosure scenarios
- Weak access protection mechanisms for ML model components
- Lack of security process to maintain a good security level of the components of the ML application
- Unprotected sensitive data on test environments
- Too much information about the model given in its outputs
- The model can allow private information to be retrieved
- Disclosure of sensitive data for ML algorithm training
- Too much information available on the model
- Too much information about the model given in its outputs

Compromise of ML application components - Definition

- Compromise of a component or developing tool of the ML application.
  - Example: compromise of one of the open-source libraries used by the developers to implement the ML algorithm.

Compromise of ML application components - Vulnerabilities

- Poor access rights management
- Too much information available on the model
- Existence of several vulnerabilities because the ML application was not included into process for integrating security into projects
- Use of vulnerable components (among the whole supply chain)
- Too much information about the model given in its outputs
- Existence of unidentified compromise scenarios
- Undefined indicators of proper functioning, making complex compromise identification
- Bad practices due to a lack of cybersecurity awareness
- Lack of security process to maintain a good security level of the components of the ML application

Failure or malfunction of ML application - Definition

- This threat refers to ML application failure (e.g. denial of service due to bad input, unavailability due to a handling error).
- The different stakeholders of the model can make mistakes that result in a failure or malfunction of ML application (human error).
  - Example: due to lack of documentation, they may use the application in use-cases not initially foreseen.
- Input data whose format is inappropriate (denial of service).
  - Example: a malicious user of the model constructs an input data (a sponge example) specifically designed to increase the computation time of the model and thus potentially cause a denial of service.

Failure or malfunction of ML application - Vulnerabilities

- Existing biases in the ML model or in the data
- ML application not integrated in the cyber-resilience strategy
- Existence of unidentified failure scenarios
- Undefined indicators of proper functioning, making complex malfunction identification
- Lack of explainability and traceability of decisions taken
- Lack of security process to maintain a good security level of the components of the ML application
- Existence of several vulnerabilities because ML specificities are not integrated in existing policies
- Contract with a low security third party

Security Controls (Countermeasures)

- ENISA identified a list of 37 security controls
- Organisational and Policy
  - More traditional security controls, either organisational or linked to security policies.
- Technical
  - More classic technical security controls.
- **Specific to ML**
  - **Security controls that are specific to applications using ML.**

Organisational

- Apply a RBAC model, respecting the least privileged principle
    - RBAC = Role Based Access Control
- Apply documentation requirements to AI projects
- Assess the regulations and laws the ML application must comply with
- Ensure ML applications comply with data security requirements
- Ensure ML applications comply with identity management, authentication, and access control policies
- Ensure ML applications comply with security and protection policies and are integrated to security operations processes
- Include ML applications into detection and response to security incident processes
- Include ML applications in asset management processes

Technical

- Assess the exposure level of the model used
- Check the vulnerabilities of the components used so that they have an appropriate security level
- Conduct a risk analysis of the ML application
- Control all data used by the ML model
- Define and monitor indicators for proper functioning of the model
- Ensure appropriate protection is deployed for test environments
- Ensure ML applications comply with third parties' security requirements
- Ensure ML projects follow the global process for integrating security into projects

Specific to ML (1/2)

- Add some adversarial examples to the training dataset
- Apply modifications on inputs
- Build explainable models
- Choose and define a more resilient model design
- Enlarge the training dataset
- Ensure that models are unbiased
- Ensure that models respect differential privacy to a sufficient degree
- Ensure that the model is sufficiently resilient to the environment in which it will operate.
- Implement processes to maintain security levels of ML components over time

Specific to ML (2/2)

- Implement tools to detect if a data point is an adversarial example or not
- Integrate ML specificities to awareness strategy and ensure all ML stakeholders are receiving it
- Integrate poisoning control after the "model evaluation" phase
- Use less easily transferable models
- Reduce the available information about the model
- Reduce the information given by the model
- Use federated learning to minimize risk of data breaches
- Use less easily transferable models

Leassons learned…so far (2022)

- There is no silver bullet for mitigating ML-specific attacks
    - Some security controls may be bypassed by adaptive attackers.
    - Applied mitigations can still raise the bar for attackers.
- ML-specific mitigation controls are not generally evaluated in a standardised way even if it is a current and important issue to enable comparability.
    - More research should be devoted to standardised benchmarks for comparing ML-specific mitigations on a level playing field.
    - These benchmarks should also be enforced to ensure that the methods used in practice are the ones that perform best.
- Deploying security controls often leads to a trade-off between security and performance
    - This is a topic of particular importance that should be further pursued by the research and cybersecurity communities.

Exercise:

1) Find potential vulnerabilities of the proposed AI-based architecture and its associated threats
2) Propose countermeasures, link them with ENISA security controls

M3G.1.pdf

OFC 2023 © 2023 Optica Publishing Group

An Intelligent Optical Telemetry Architecture

Luis Velasco*, Pol González, and Marc Ruiz

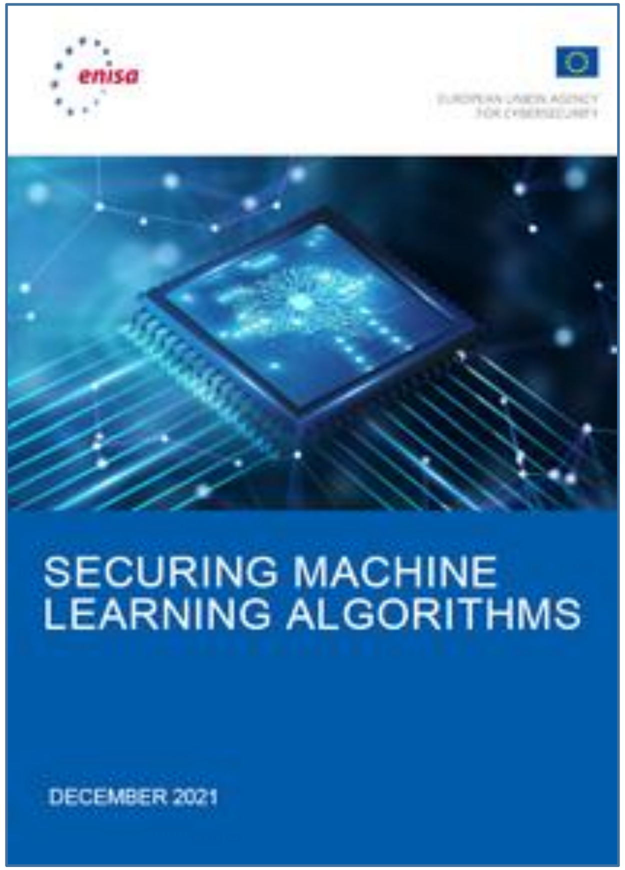
Optical Communications Group (GCO), Universitat Politècnica de Catalunya (UPC), Barcelona, Spain
e-mail: luis.velasco@upc.edu

Abstract: A distributed telemetry system is proposed with agents receiving and analyzing data before sending to a centralized manager. Intelligent data aggregation on optical constellations telemetry largely reduces data rate without introducing significant error.

Link to the document

References

enisa
EUROPEAN
UNION AGENCY
FOR CYBERSECURITY



Cybersecurity Management
GCS 2.3 – AI/ML

2022-2023

Prof. Marc Ruiz

marc.ruiz-ramirez@upc.edu