OWASP™

THALES
Building a future we can all trust

# Safeguarding ML: A comprehensive security plan

**Viswanath S CHIRRAVURI**

**October 2024**

www.thalesgroup.com

# Artificial Intelligence… What are we talking about?

## Artificial Intelligence

Any technique that enables computers to mimic human intelligence

## Machine Learning

A subset of Artificial Intelligence focusing on data-based learning by opposition to symbolic AI (aka knowledge-based learning).

## Deep Learning

A subset of Machine Learning methods, based on Artificial Neural Networks with many layers (aka Deep) Neural Networks). There are many types of ANN (e.g. CNN: Convolutional Neural Network, RNN: Recurrent Neural Network).

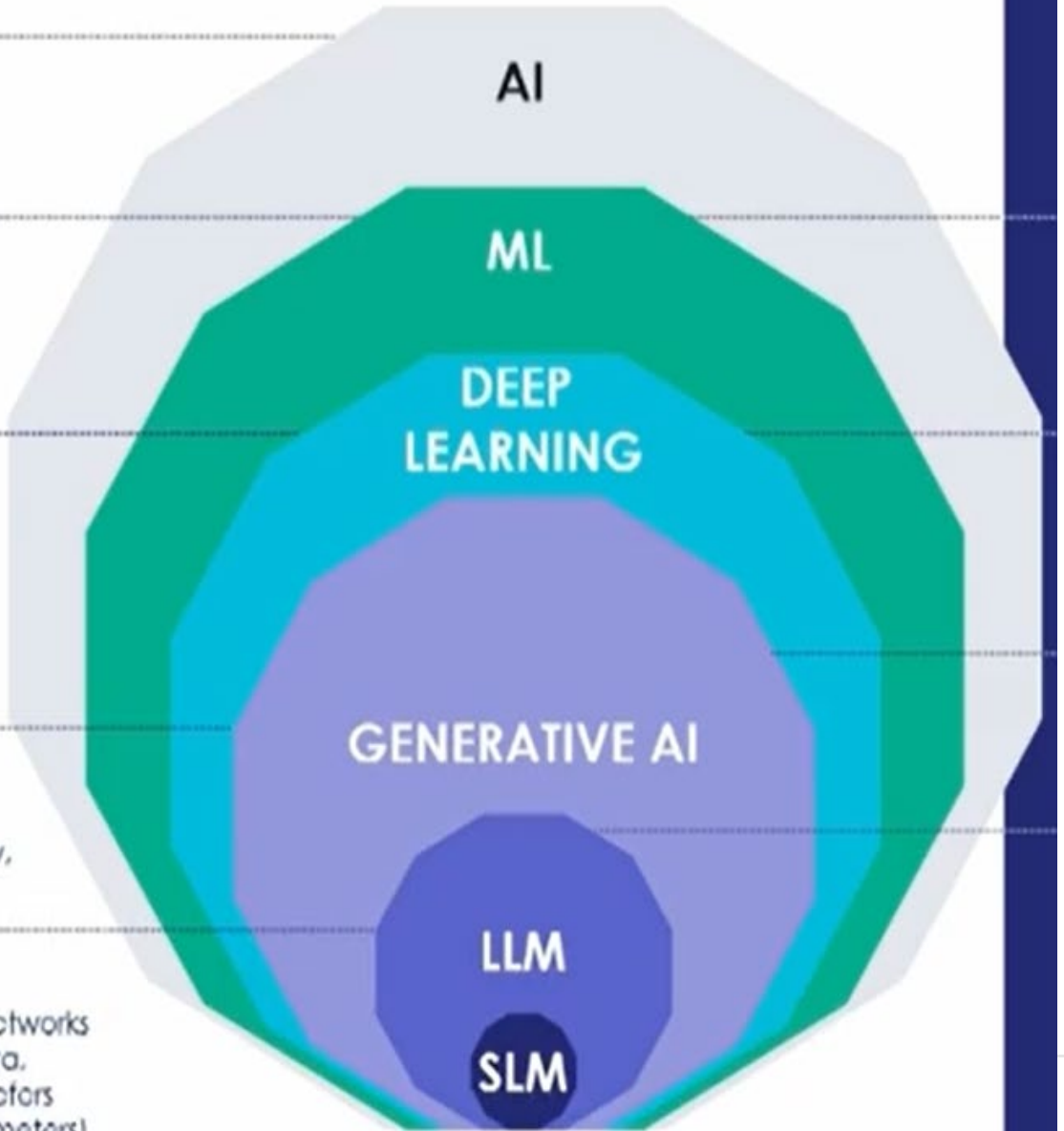
## Generative AI

A type of artificial intelligence technology that can produce various types of content, including text, imagery, audio and synthetic data. Examples: GAN*, LLM…

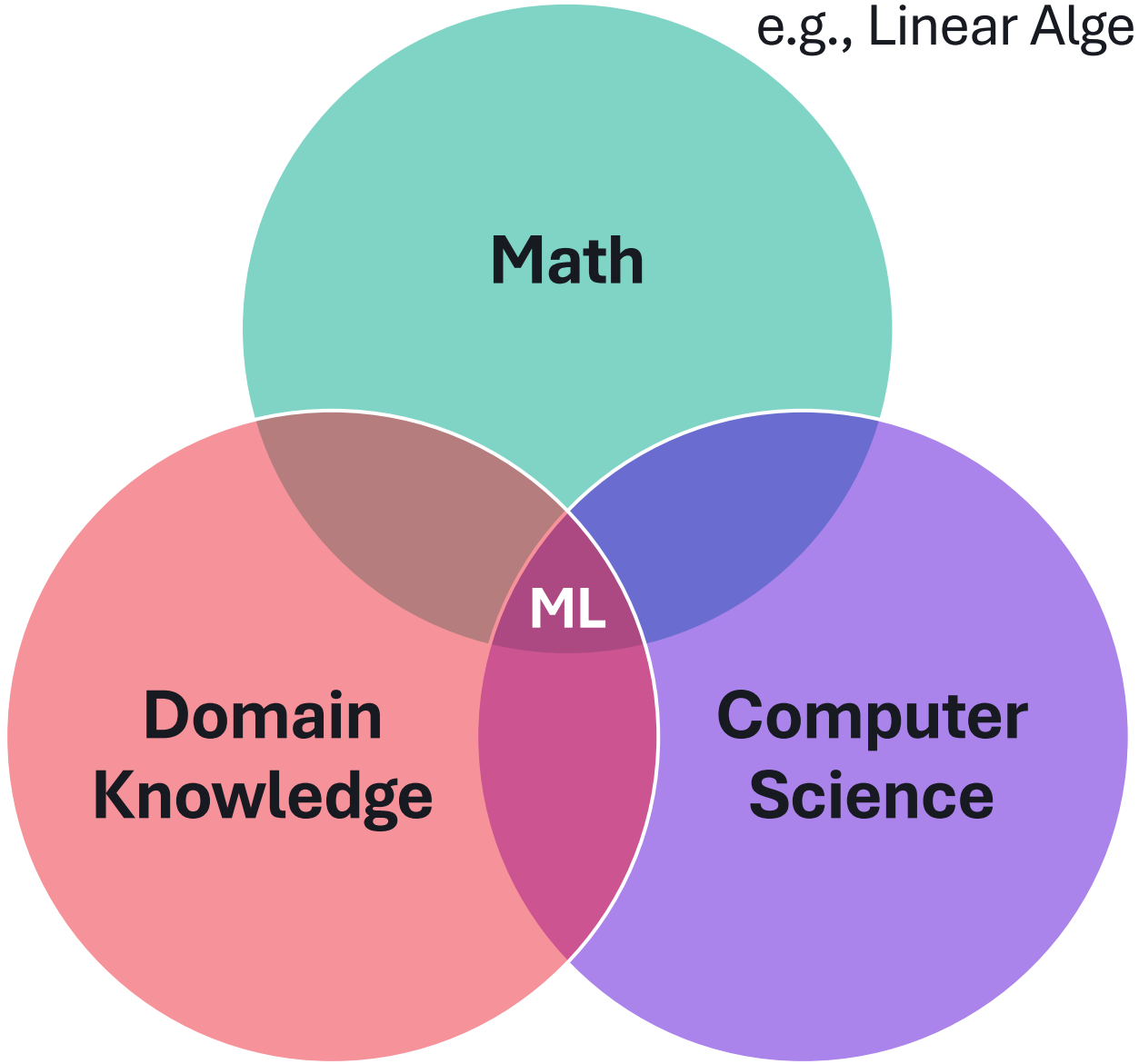
## Large Language Models

Specific application of GenAI using a variety of neural networks (the transformers), trained on a very large amount of data, commonly coming with XX+ billions of parameters. SLM refers to Small Language Model (in millions or X billions of parameters).

*GAN: Generative Adversarial Network



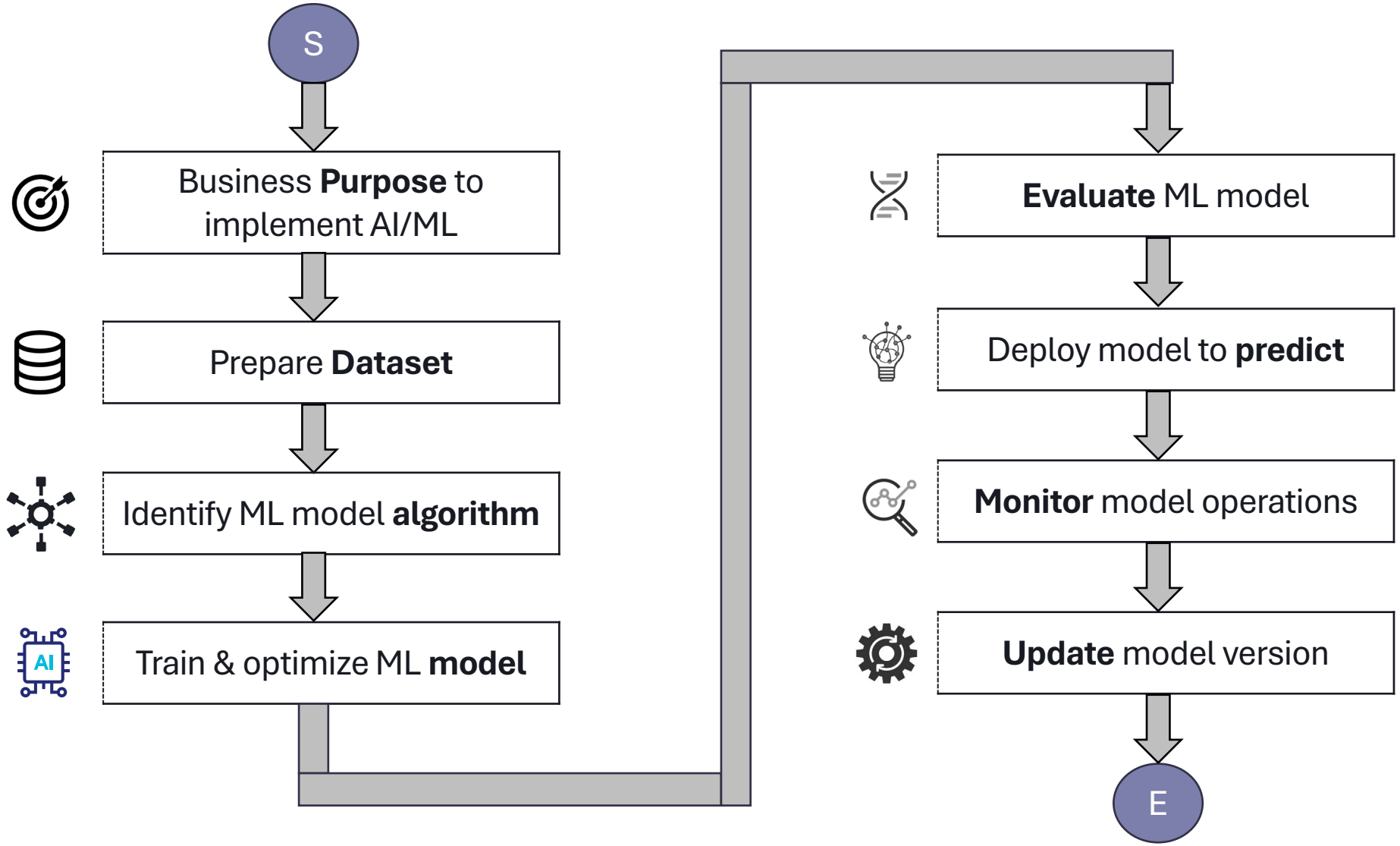
e.g., Linear Algebra



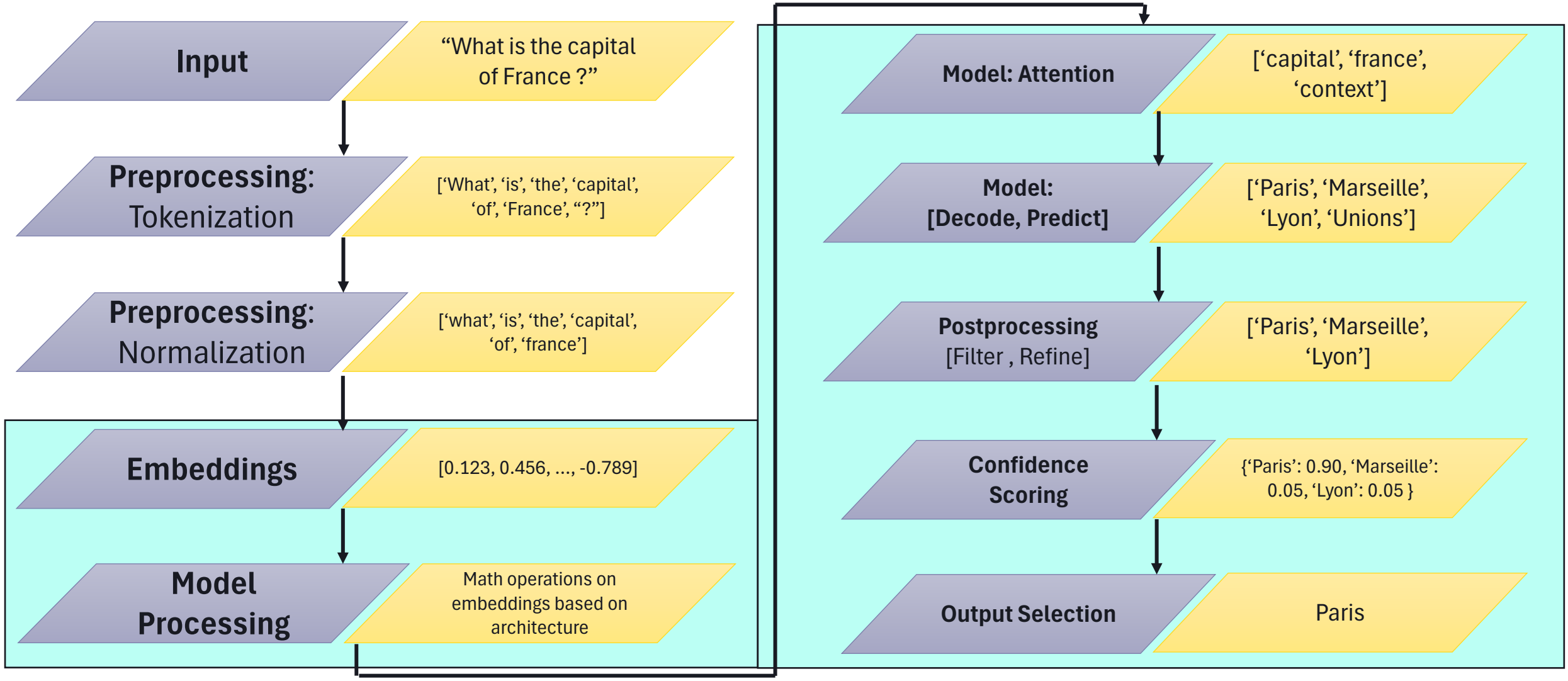

e.g., Cybersecurity

e.g., Python

# Machine Learning Workflow



# What happens when you input something into LLMs in production? (with example)



# Artificial Intelligence and Cybersecurity

## AI for Cybersecurity

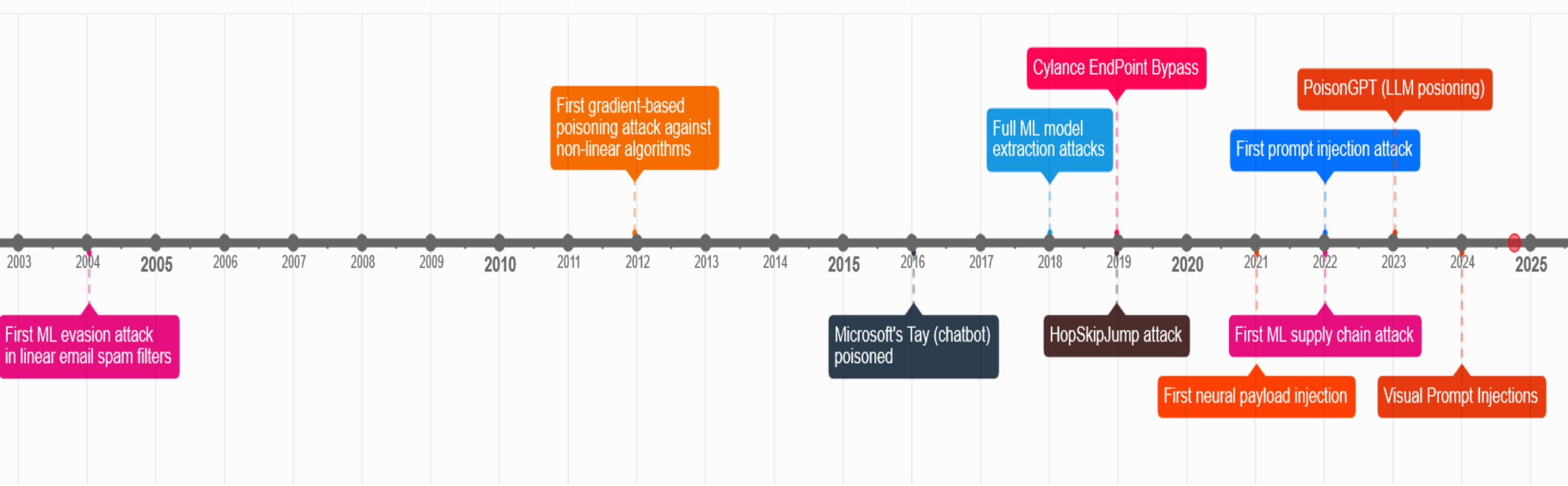
- Implementing AI/ML to protect systems (defense) e.g., AI for improved Data Protection
- Implementing AI/ML to hack systems (attack) e.g., AI generated malware, AI to scan for vulns
- Implementing AI/ML in forensics or incident handling (resolve) e.g., AI for threat hunting
- Implementing AI/ML in management (report) e.g., AI for Cybersecurity metrics / KPIs

## Cybersecurity for AI

- Implementing security in AI/ML business usecases (Trusted AI/ML) e.g., Securing LLMs

# Evolution of Cyberattacks in AI/ML

Source: HiddenLayer

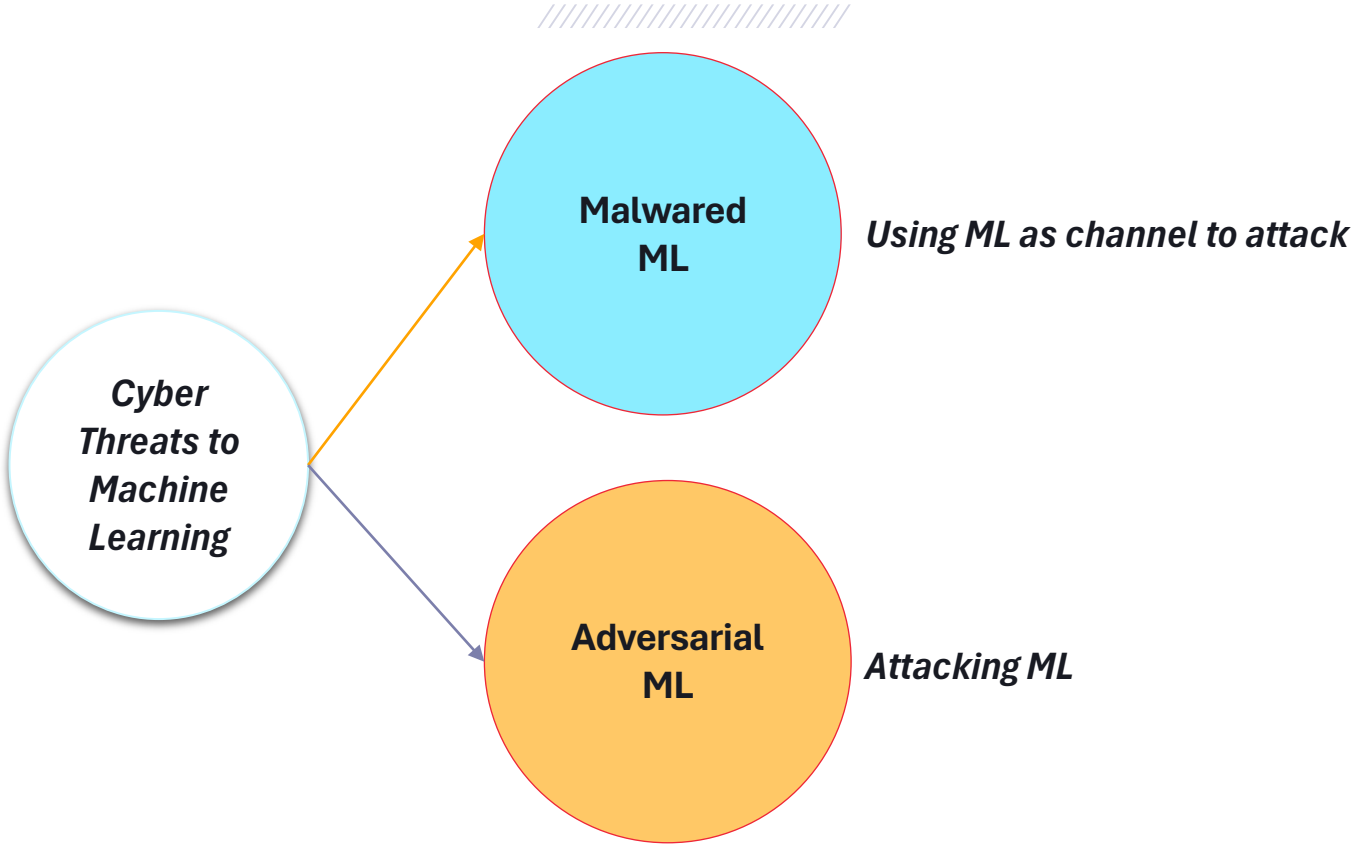



https://oecd.ai/en/incidents
https://airisk.io/
https://avidml.org/database/
https://incidentdatabase.ai/
https://atlas.mitre.org/studies

# BUSINESS RISKS TO IMPLEMENT AI / ML

- Harmful Content Creation
- Deepfakes
- Copyright Violation
- Reputational Damage
- Increased Costs
- Regulatory Fines
- Intellectual Property Loss
- Business Continuity Disruption
- Customer Loss

# CYBER THREAT CATEGORIES TO MACHINE LEARNING



# Cyber Threats to Machine Learning

## Malwared ML

ML datasets, pre-trained models, or underlying ML libraries like TensorFlow, PyTorch, scikit-learn, etc. containing **malware** to spread across networks.

- Virus
- Worms
- Trojan horse
- Spyware
- Adware
- Botnets
- Ransomware

## Adversarial ML

Manipulate ML datasets, models, or systems with the goal of compromising the confidentiality, integrity, or availability of the business use case.

- Model Evasion
- Model Extraction / Theft
- Model Backdoor
- Model Denial of service
- Data Poisoning / Data Backdoor
- Data Theft / Model Inference
- Generative AI: Prompt / Code Injection / Jailbreaking

# AI/ML: Regulations, Standards, Frameworks

## Regulations

- US White House EO 14110
- EU AI Act
- Canada AI & Data Act
- UK AI Bill
- PDPC Singapore

## Standards

- NIST AI RMF
- ENISA & ETSI
- OWASP Top 10 (ML & LLM)
- ISO 27090
- IEEE P3119

## Frameworks

- MITRE ATLAS
- Gartner AI TRiSM
- Linux Foundation AIOSS
- Google SAIF
- IBM Secure Gen AI
- Thales Secure ML

# ThalesGroup's **Secure Machine Learning** Framework

- We open sourced this secure-machine-learning framework
  - https://github.com/ThalesGroup/secure-ml [License: CC BY-ND 4.0]

- Includes:
  - Corporate Security Policy: Framework for ML systems [Link]
  - Security Requirements & Guidelines [Link]
  - Privacy-Preserving Techniques [Link]
  - Recommended Security Tools [Link]
  - Industry references on regulations, standards [Link]
  - Cyber threats to ML & Taxonomy [Link]

# **ML Security Policy** – highlights


- **Data Security**: It refers to securing data that can be training data, validation data, data stored in vector databases, etc.
- **Model Security**: It refers to hardening the model in both development and inference time.
- **Platform Security**: It refers to hardening of the underlying platform where the ML model is being developed or operated, along with any associated data.
- **Security Compliance**: It refers to being compliant to both internal and external regulations, standards, and best practices as application to the Trusted AI principles.
- **Human Security**: It refers to humans being well trained, authenticated, authorized, and capable to handle security incidents around ML ecosystem.

| | DATA SECURITY | MODEL SECURITY | PLATFORM SECURITY | SECURITY COMPLIANCE | HUMAN SECURITY |
|---|---|---|---|---|---|
| **GOAL** | • Confidentiality<br>• Integrity<br>• Availability<br>• Authentication<br>• Authorization<br>• Non-repudiation<br>• Privacy | • Integrity in computation<br>• Accuracy and precision in output | • API security<br>• System security<br>• Network security | • Comply with internal and external policies, standards and regulations | • People involved are aware of security risks<br>• Only authorized people are involved |
| **TECHNIQUE** | • Encryption<br>• Access Controls<br>• Data backups & Recovery<br>• Data anonymization<br>• Data quality control<br>• Secure data sharing<br>• Data classification | • Secure development<br>• Input and Output validation<br>• Model explainability<br>• Adversarial training<br>• Robustness testing<br>• Monitoring and alerting<br>• Secure deployment | • Vulnerability scanning & Penetration Testing<br>• Patch management<br>• Access Controls<br>• Encryption<br>• Hardening<br>• Secure Configuration | • Regulatory compliance<br>• Ethical considerations<br>• Data retention and deletion<br>• Audit trails<br>• Security assessments<br>• Third-party risk management | • Training and awareness<br>• Background checks<br>• Incident response<br>• Governance and oversight<br>• Continuous monitoring |

# ML Security Requirements – DATA SECURITY

WHAT?

## Data Security Techniques

| ID | Technique | Activity |
|---|---|---|
| DS1 | Encryption | - Encrypt sensitive data in transit and at rest.<br>- Use strong algorithms (e.g., AES, RSA).<br>- Securely manage encryption keys. |
| DS2 | Access Controls | - Implement strict access controls.<br>- Ensure authorized personnel access only.<br>- Utilize user authentication, role-based access, and MFA. |
| DS3 | Data Classification, Backups, and Recovery | - Classify data by sensitivity.<br>- Regularly back up data.<br>- Develop a disaster recovery plan. |
| DS4 | Data Anonymization | - Remove personally identifiable information (PII).<br>- Use data masking or tokenization techniques to protect privacy. |
| DS5 | Data Quality Control | - Implement processes for data quality and integrity.<br>- Conduct data validation and antivirus scans. |
| DS6 | Secure Data Sharing | - Ensure security measures for third-party data sharing.<br>- Use secure transfer protocols and establish data usage agreements. |

## Key Takeaways

1. Implement Robust Security Measures: Focus on encryption, access controls, and data quality.
2. Protect Sensitive Information: Anonymization and secure sharing are essential.
3. Plan for Recovery: Regular backups and disaster recovery plans ensure data integrity.

# ML Security Requirements – MODEL SECURITY

WHAT?

## Model Security Techniques

| ID | Technique | Activity |
|---|---|---|
| MS1 | **Secure Development** | - Implement version control for AI models.<br>- Use code repositories and branching.<br>- Conduct code reviews and maintain a secure development environment. |
| MS2 | **Input/Output Validation** | - Validate inputs and outputs for accuracy and reliability.<br>- Employ data validation, normalization, and transformation to eliminate untrusted inputs. |
| MS3 | **Model Explainability** | - Ensure AI models are explainable and transparent.<br>- Utilize feature importance analysis and model visualization.<br>- Conduct threat modeling. |
| MS4 | **Adversarial Training & Robustness Testing** | - Perform robustness testing against malware and adversarial attacks.<br>- Implement scanning, adversarial training, and model hardening techniques. |
| MS5 | **Monitoring and Alerting** | - Monitor AI models in production for anomalies.<br>- Set up logging and alerting mechanisms for unexpected behavior. |
| MS6 | **Secure Deployment** | - Ensure secure deployment of AI models.<br>- Use secure communication protocols and access controls.<br>- Implement authentication and reverse engineering protection. |

## Key Takeaways

1. Integrity and Resilience: Focus on secure development and robustness against attacks.
2. Transparency: Prioritize model explainability for better understanding and trust.
3. Continuous Monitoring: Regularly monitor models to detect anomalies and ensure secure deployment.

# ML Security Requirements – PLATFORM SECURITY

WHAT?

## Platform Security Techniques

| ID | Technique | Activity |
|---|---|---|
| PS1 | **Vulnerability Scanning** | - Perform vulnerability scanning and penetration testing.<br>- Identify and mitigate vulnerabilities in AI/ML infrastructure and systems. |
| PS2 | **Patch Management** | - Implement a robust patch management process.<br>- Ensure timely installation of software and firmware updates to address known vulnerabilities. |
| PS3 | **Access Controls** | - Enforce strict access controls for AI/ML infrastructure.<br>- Use user authentication, role-based access control, and multi-factor authentication. |
| PS4 | **Encryption** | - Utilize encryption for data in transit and at rest.<br>- Implement TLS for secure communications and full-disk encryption for storage. |
| PS5 | **Network Hardening** | - Apply network security measures such as firewalls and intrusion detection/prevention systems.<br>- Implement network segmentation to defend against attacks. |
| PS6 | **Hardware Security** | - Secure hardware components (e.g., GPUs, TPUs) against tampering and physical attacks. |
| PS7 | **Secure Configuration** | - Ensure secure configuration of ML systems.<br>- Follow best practices for hardening networks, systems, and applications, both in cloud and on-premises. |

## Key Takeaways

1. Proactive Security: Regular vulnerability scanning and patch management are essential.
2. Controlled Access: Implement robust access controls to safeguard infrastructure.
3. Comprehensive Protection: Ensure both hardware and software configurations are secure.

# ML Security Requirements – SECURITY COMPLIANCE

WHAT?

## Security Compliance Techniques

| ID | Technique | Activity |
|---|---|---|
| SC1 | **Regulatory Compliance** | - Ensure AI/ML systems comply with regulations (e.g., HIPAA, GDPR, PCI-DSS, ISO 27001).<br>- Implement data protection, security controls, and privacy requirements. |
| SC2 | **Ethical Considerations** | - Consider ethical implications, focusing on bias, fairness, and accountability.<br>- Employ bias mitigation techniques and promote algorithmic transparency. |
| SC3 | **Data Retention and Deletion** | - Establish policies for data retention and deletion to meet regulatory requirements.<br>- Implement data minimization, anonymization, and destruction practices. |
| SC4 | **Audit Trails** | - Create audit trails to track data access and usage.<br>- Utilize logging, monitoring, and reporting, especially in regulated industries. |
| SC5 | **Security Assessments** | - Conduct security assessments to identify vulnerabilities and ensure compliance.<br>- Perform security risk assessments, audits, and compliance evaluations. |
| SC6 | **Third-Party Risk Management** | - Ensure third-party vendors comply with regulatory requirements and security standards.<br>- Assess risks associated with vendors accessing sensitive data or systems. |

## Key Takeaways

1. Compliance Matters: Adhere to regulations to protect data and privacy.
2. Ethical Frameworks: Address ethical considerations to ensure fairness and accountability.
3. Thorough Assessments: Regularly assess security and third-party risks for robust compliance.

OPEN

# ML Security Requirements – HUMAN SECURITY

WHAT?

## Human Security Techniques

| ID | Technique | Activity |
|---|---|---|
| HS1 | **Training and Awareness** | - Provide training programs for employees and stakeholders.<br>- Implement security awareness campaigns and conduct phishing simulations to highlight risks. |
| HS2 | **Background Checks** | - Conduct background checks on employees and contractors with access to sensitive data.<br>- Focus on regulated industries and government agencies. |
| HS3 | **Incident Response** | - Develop an incident response plan for handling security incidents.<br>- Include incident management, forensics, and communication plans with designated personnel. |
| HS4 | **Governance and Oversight** | - Establish governance structures for secure and ethical AI/ML system development and deployment.<br>- Implement compliance frameworks and security dashboards. |

## Key Takeaways

1. Empower Employees: Training and awareness are crucial for understanding security risks.
2. Vetting Access: Background checks enhance security for sensitive data access.
3. Preparedness: An effective incident response plan ensures readiness for security incidents.

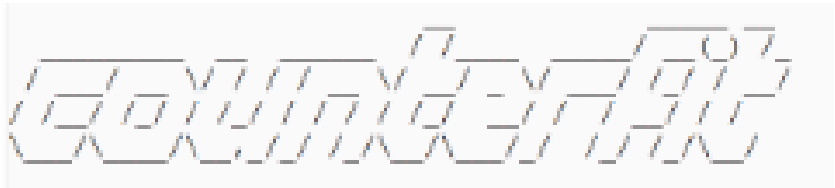
HOW?

# ML Security Guidelines

| | GUIDELINES |
|---|---|
| **DATA SECURITY** | OWASP ASVS, ISO27001, GDPR, IEEE, + custom guidelines |
| **MODEL SECURITY** | SHAP/LIME, ProtectAI' s LLM-guard, Microsoft AI Threat modeling + custom guidelines |
| **PLATFORM SECURITY** | Custom guidelines |
| **SECURITY COMPLIANCE** | FHE & FL (w/ multiple libraries), + custom guidelines |
| **HUMAN SECURITY** | Custom guidelines |

# Security tools for Machine Learning

Azure/PyRIT






QData/TextAttack

# Privacy-Preserving Techniques for Machine Learning

## Differential Privacy for Machine Learning

- https://github.com/OpenMined/PyDP
- https://github.com/tensorflow/privacy
- https://github.com/opendp/opendp

## Federated Learning

- https://github.com/tensorflow/federated
- https://github.com/OpenMined/PySyft
- https://flower.dev/

## Homomorphic Encryption

- https://github.com/OpenMined/TenSEAL
- https://github.com/zama-ai/concrete-ml
- SEAL (Simple Encrypted Arithmetic Library): https://www.microsoft.com/en-us/research/project/microsoft-seal/
- https://github.com/homenc/HElib
- More at https://github.com/jonaschn/awesome-he

## Secure Multi-Party Computation (SMPC)

- https://github.com/data61/MP-SPDZ
- https://sharemind.cyber.ee/sharemind-mpc/
- https://github.com/easy-smpc/easy-smpc

## Privacy-Preserving Data Synthesis

- https://github.com/pytorch/opacus
- https://github.com/sdv-dev/SDV
- https://github.com/sdv-dev/CTGAN
- https://github.com/DataResponsibly/DataSynthesizer

# Secure ML Program Roadmap: 3, 6, 12-Month Plan

## 3

### Foundation

1. Draft/refine corporate security policy for ML systems.
2. Plan/red team AI/ML systems (create team if needed).
3. Analyze security regulations and standards for AI/ML
4. Raise awareness among ML engineers on security risks.

## 6

### Build

1. Research security tools for your ML tech stack.
2. Build guidelines for securing data pipelines and models.
3. Explore privacy-preserving techniques (e.g., FHE, FL).
4. Begin threat modeling for ML risks.

## 12

### Launch

1. Deliver secure ML solutions to clients.
2. Implement continuous monitoring for threats.
3. Automate security compliance for ML datasets & models.
4. Train stakeholders on new ML attack vectors.

# Learn by hands-on exercises

1. **Exploring Adversarial Machine Learning** by NVIDIA

- https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-DS-03+V1

2. **Adversarial Machine Learning**

- https://secml.readthedocs.io/en/v0.15/tutorials.adv.html

3. **Dreadnode's crucible (CTF)**

- https://crucible.dreadnode.io/

4. **Lakera's Gandalf (8-levels) for prompt injection attack (CTF)**

- https://gandalf.lakera.ai/baseline

LASCON

OWASP

THALES
Building a future we can all trust

**Viswanath S CHIRRAVURI**

**https://www.linkedin.com/in/chviswanath/**

# Thank you

www.thalesgroup.com