New AI and Quantum Risk and Security

Michael Melore, CISSP
National Public Sector Security Evangelist
IBM Security
Chair SecRT (Security <Leaders> Round Tables)
www.SecRT.us
@MichaelMelore
mmelore@us.ibm.com

IBM Security No Cost Trusted Advisory Programs

- **Cyber Resiliency Assessments**
  NIST Framework, vendor agnostic, non-invasive assessment
  https://www.ibm.com/storage-data-resiliency
- **Architect led collaborative workshops**
  Interactive no products pitched education
- **XForce Cyber Range Exercises**
  Exclusive Sessions available at cost
- **Solution and Threat Briefings / Studies and Reports**
  Point of contact: mmelore@us.ibm.com
  https://www.ibm.com/security
- **IBM Center for Government Cybersecurity**
  Understand cyber threats with the IBM Center for Government Cybersecurity
  https://www.ibm.com/security/services/us-federal-cybersecurity-center
- **Onsite collaborative exercises/knowledge sharing/tech days**
  Can be jointly delivered with State Officials
- **IBM Sponsored External Committees**
  www.ijis.org IJIS Institute Cybersecurity Task Force
  www.NASCIO.org National Assoc State CIOs
  Privacy and Cybersecurity Working Groups
  www.ATARC.org Advanced Technology Academic Research Center
  Government Grants, Higher Ed, and Cyber Defense Working Groups

XForce Exchange Threat Intelligence Forum
https://exchange.xforce.ibmcloud.com

IBM Security Learning Academy
Hundreds of no cost Tutorials/Hand on Labs
https://www.securitylearningacademy.com

IBM's New Collar / Apprentice programs
https://www.ibm.com/impact/feature/apprenticeship

IBM's Skills Build (Resources for Learners, Educators and Organizations)
https://skillsbuild.org

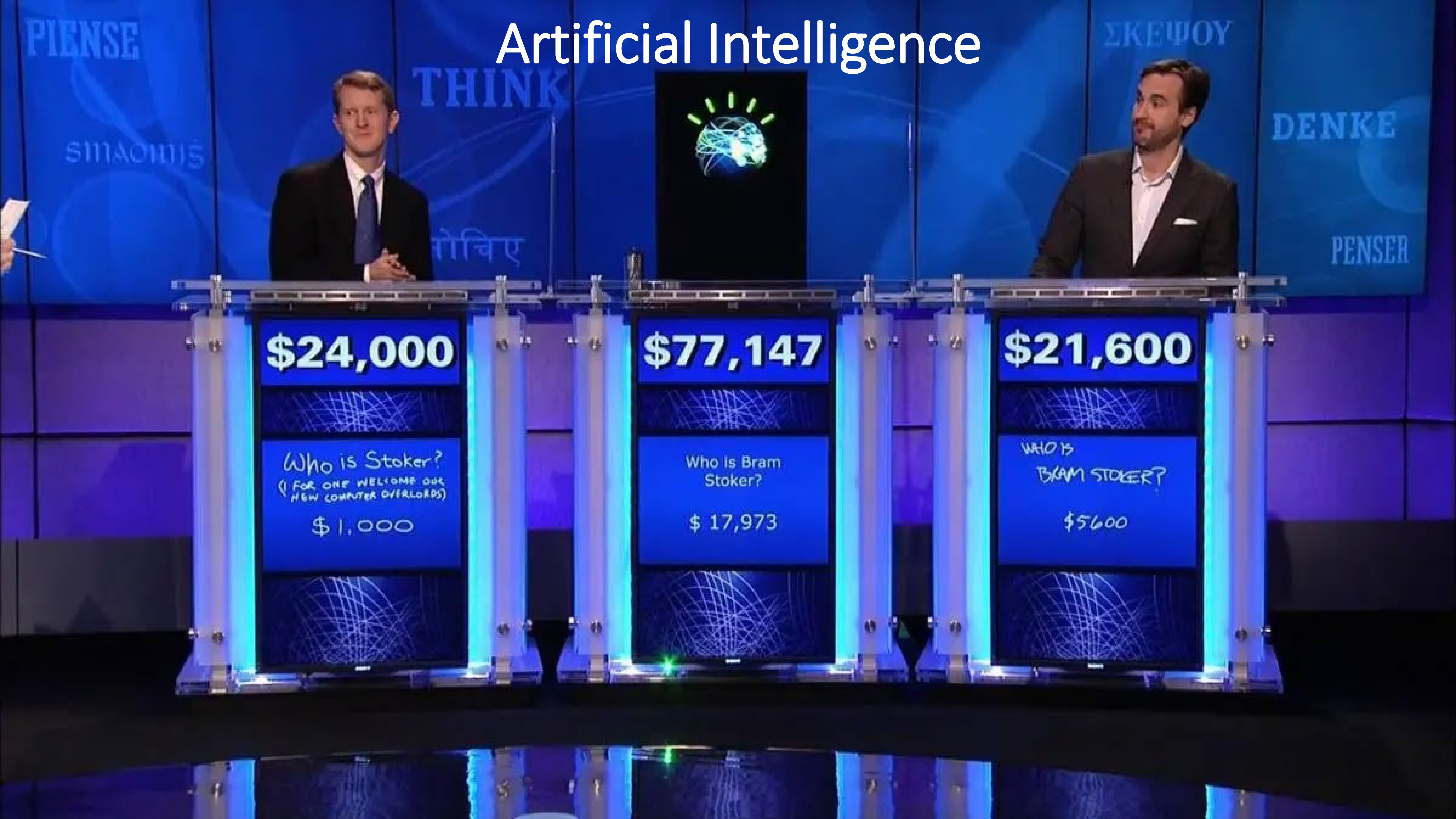
XForce Webinars on Threat Landscapes

IBM Security Solution User Groups

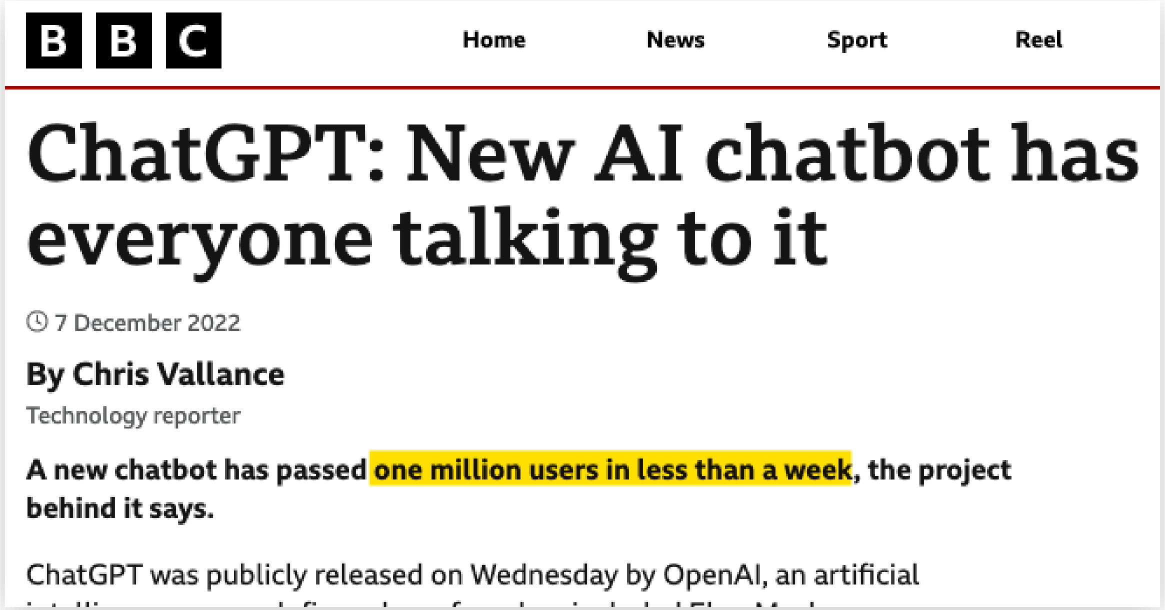
Regional/Nationwide Security Leaders Round Tables

www.SECRT.us
Grass roots, non-sponsored, no cost Security Executive Round Tables Nationwide,

Artificial Intelligence



OpenAI



ChatGPT: New AI chatbot has everyone talking to it

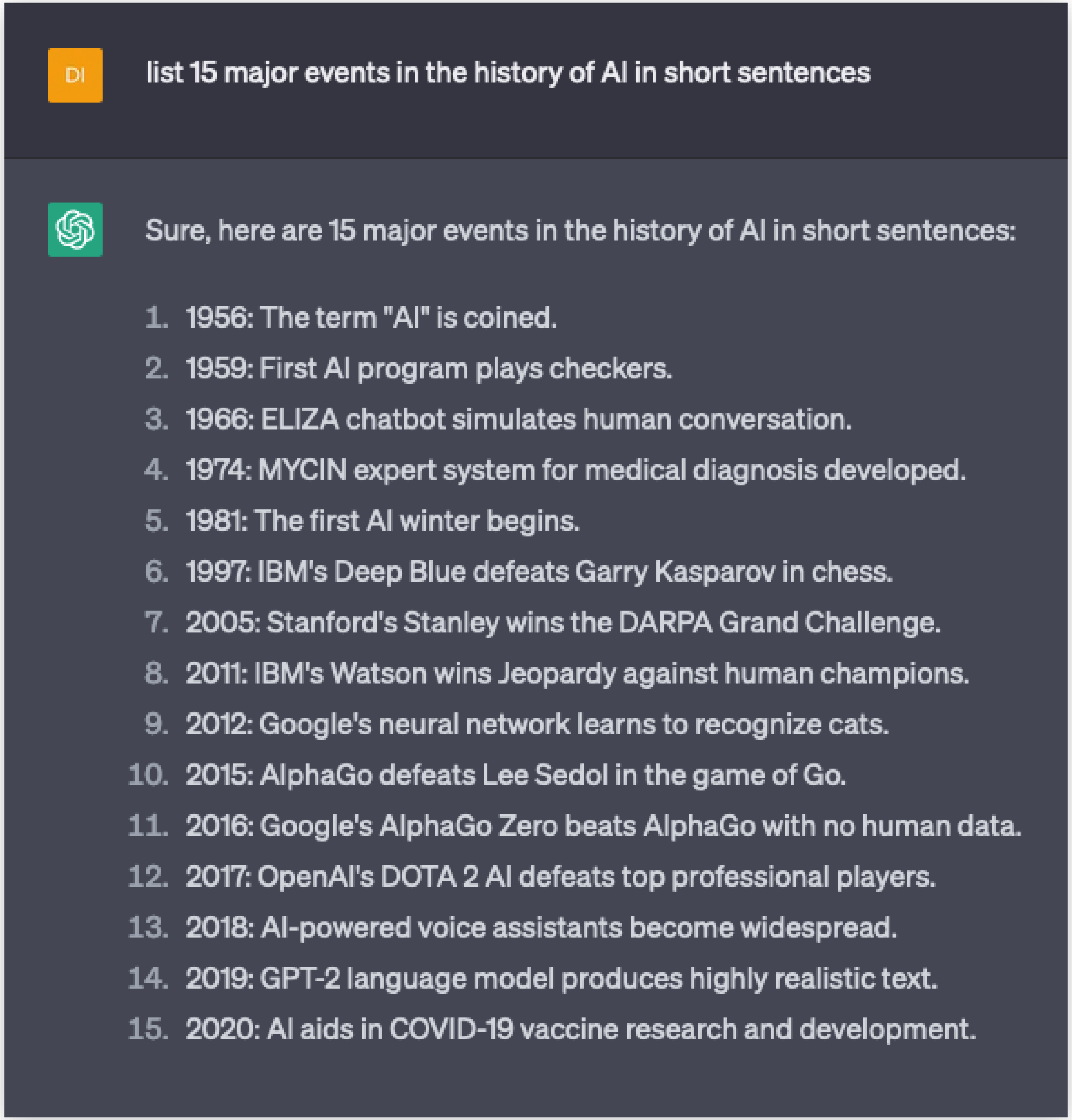
7 December 2022

By Chris Vallance

Technology reporter

A new chatbot has passed one million users in less than a week, the project behind it says.

ChatGPT was publicly released on Wednesday by OpenAI, an artificial



list 15 major events in the history of AI in short sentences

Sure, here are 15 major events in the history of AI in short sentences:

1. 1956: The term "AI" is coined.
2. 1959: First AI program plays checkers.
3. 1966: ELIZA chatbot simulates human conversation.
4. 1974: MYCIN expert system for medical diagnosis developed.
5. 1981: The first AI winter begins.
6. 1997: IBM's Deep Blue defeats Garry Kasparov in chess.
7. 2005: Stanford's Stanley wins the DARPA Grand Challenge.
8. 2011: IBM's Watson wins Jeopardy against human champions.
9. 2012: Google's neural network learns to recognize cats.
10. 2015: AlphaGo defeats Lee Sedol in the game of Go.
11. 2016: Google's AlphaGo Zero beats AlphaGo with no human data.
12. 2017: OpenAI's DOTA 2 AI defeats top professional players.
13. 2018: AI-powered voice assistants become widespread.
14. 2019: GPT-2 language model produces highly realistic text.
15. 2020: AI aids in COVID-19 vaccine research and development.

“ Interesting analysis evocative of two truths:

1. ***The first report is usually wrong***
2. ***It is exceedingly difficult to unseat a first impression “***

Heather McMahon
Artemist Advisory Group, LLC

Clearing up the Unclear

- **Misinformation** is false, but not created or shared with the intention of causing harm.
- **Disinformation** is deliberately created to mislead, harm, or manipulate a person, social group, organization, or country.
- **Malinformation** is based on fact, but used out of context to mislead, harm, or manipulate.

Business is adopting AI

AI for Business

for example…

Protect it.

- Talent
- Marketing
- Security
- Automation
- Finance
- Regulations

35% of companies are using AI in their business.

https://www.ibm.com/downloads/cas/RNZ6DYLN

Security sits between the two

Security for AI

for example…

Use it.

- Threat monitoring and response
- Privacy Controls and management
- Employee education
- Secure design and engineering
- AI Security
- Infrastructure protection

$1.76 million The effect of extensive security AI and automation on the financial impact of a breach.

https://www.ibm.com/reports/data-breach

So are attackers

Adversarial AI

for example…

Defend against it.

- Theft
- Phishing
- Social engineering
- Malware
- Fakes
- Poison

56% of identified cyberattacks leveraged AI in the access and penetration phase.

https://eftsure.com/statistics/artificial-intelligence-statistics/#section-26

Attacker’s Use of AI in Security

Lightboard talk on Chatbot poisoning
https://youtu.be/RTCaGwxD2uU

AI Powered Attacks

- **Generate:** DeepHack tool learned SQL injection
- **Automate:** Generate targeted phishing attacks on Twitter
- **Refine:** Neural network powered password crackers
- **Evade:** Generative adversarial networks learn novel steganographic channels

Attacking AI

- **Poison:** Microsoft Tay chatbot poisoning via Twitter (and Watson Urban Dictionary “poisoning”)
- **Evade:** Real-world attacks on computer vision for facial recognition biometrics and autonomous vehicles
- **Harden:** Genetic algorithms and reinforcement learning (OpenAI Gym) to evade malware detectors

Theft of AI

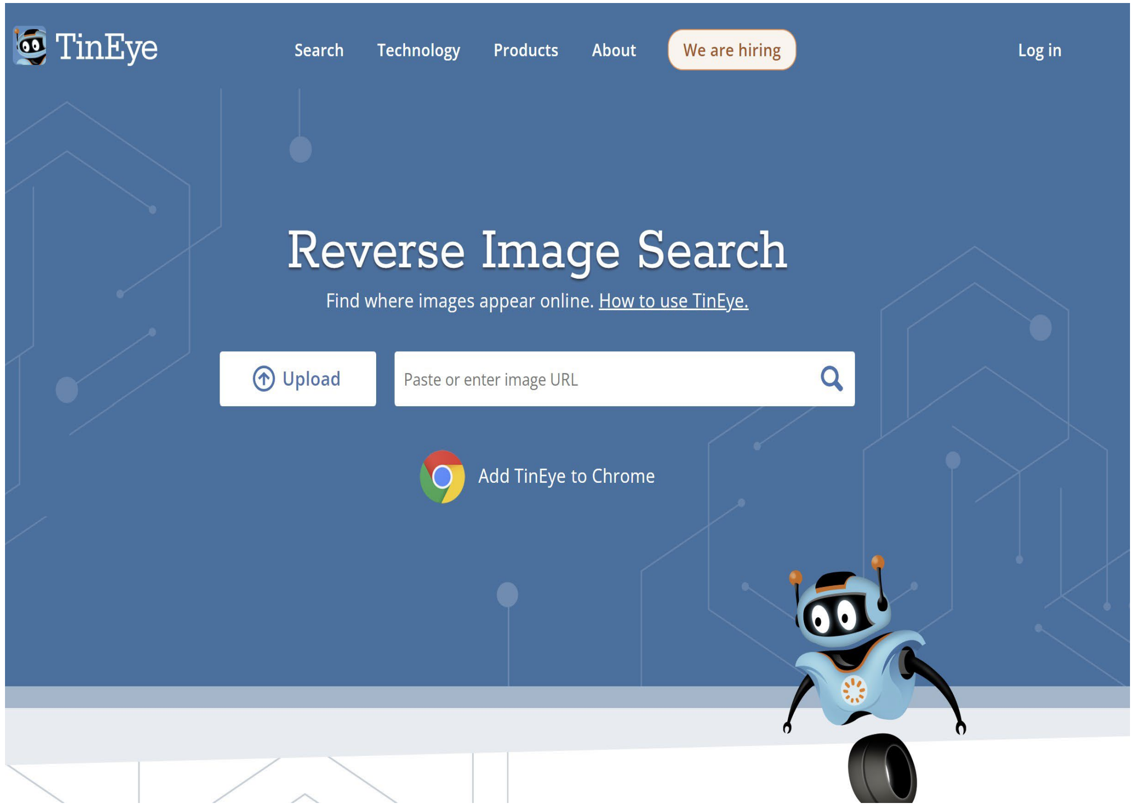
- **Theft:** Stealing machine learning models via public APIs
- **Transferability:** Practical black-box attacks learn surrogate models for transfer attacks
- **Privacy:** Model inversion attacks steal training data

Image Source: DALL·E 2 | OpenAI

Tools to help find Deep Fakes

TinEye Reverse Image Search

https://tineye.com/

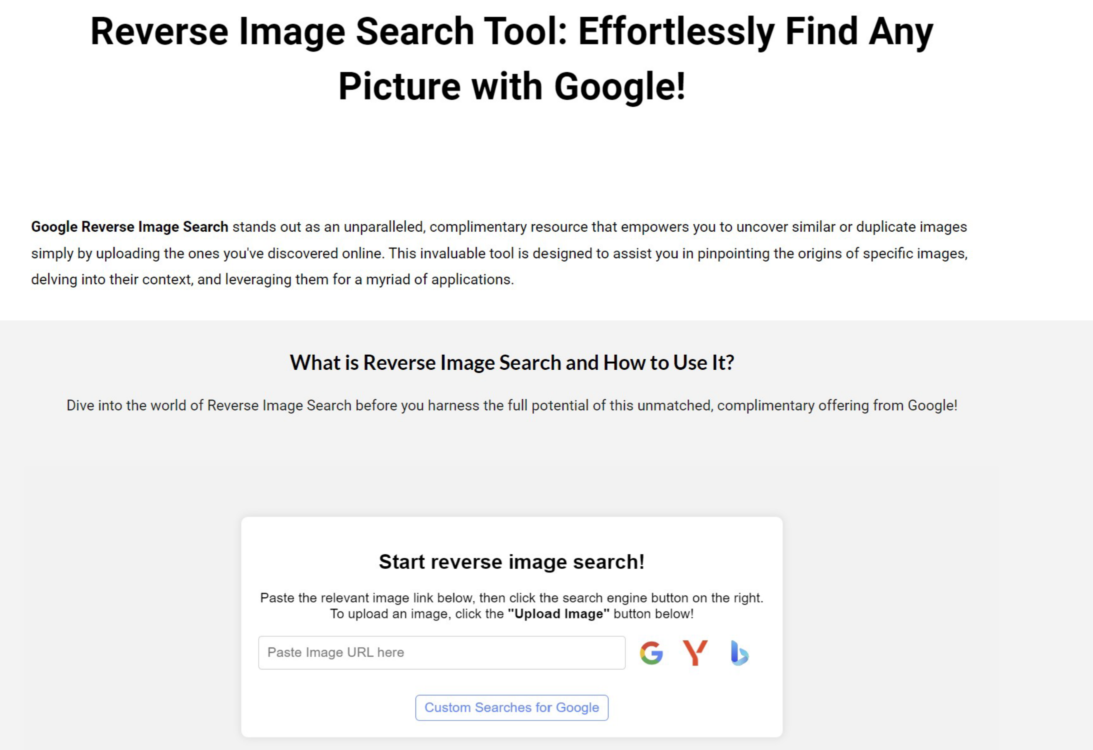


Reverse Image Search

Find where images appear online. How to use TinEye.

Google Reverse Image Search

https://sites.google.com/view/reverse-images/home



Reverse Image Search Tool: Effortlessly Find Any Picture with Google!

Google Reverse Image Search stands out as an unparalleled, complimentary resource that empowers you to uncover similar or duplicate images simply by uploading the ones you've discovered online. This invaluable tool is designed to assist you in pinpointing the origins of specific images, delving into their context, and leveraging them for a myriad of applications.

What is Reverse Image Search and How to Use It?

Dive into the world of Reverse Image Search before you harness the full potential of this unmatched, complimentary offering from Google!

Start reverse image search!

Paste the relevant image link below, then click the search engine button on the right. To upload an image, click the **"Upload Image"** button below!

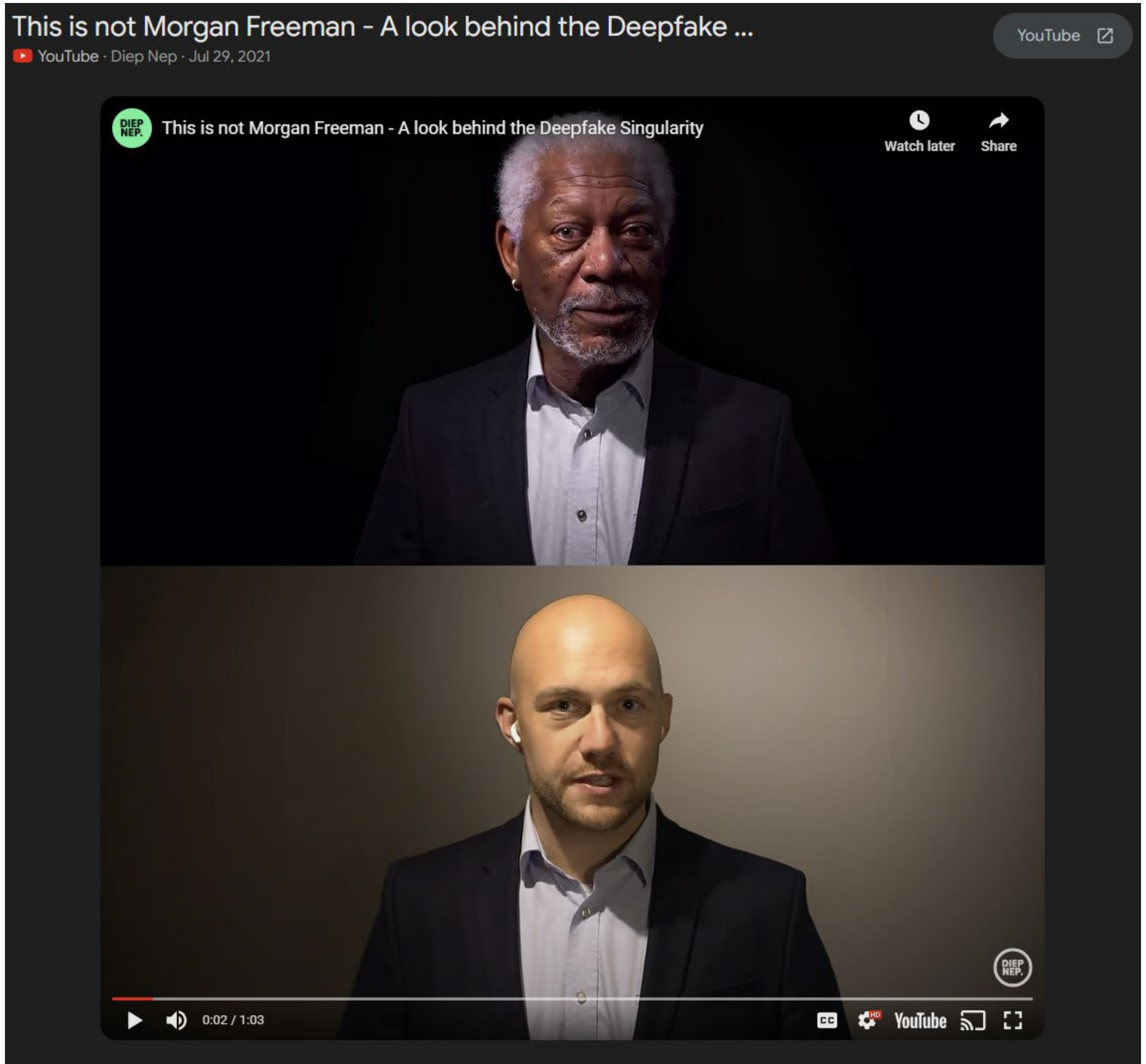
GANs – Generative Adversarial Networks

Generator and Discriminator

Source: [DALL·E 2 | OpenAI]

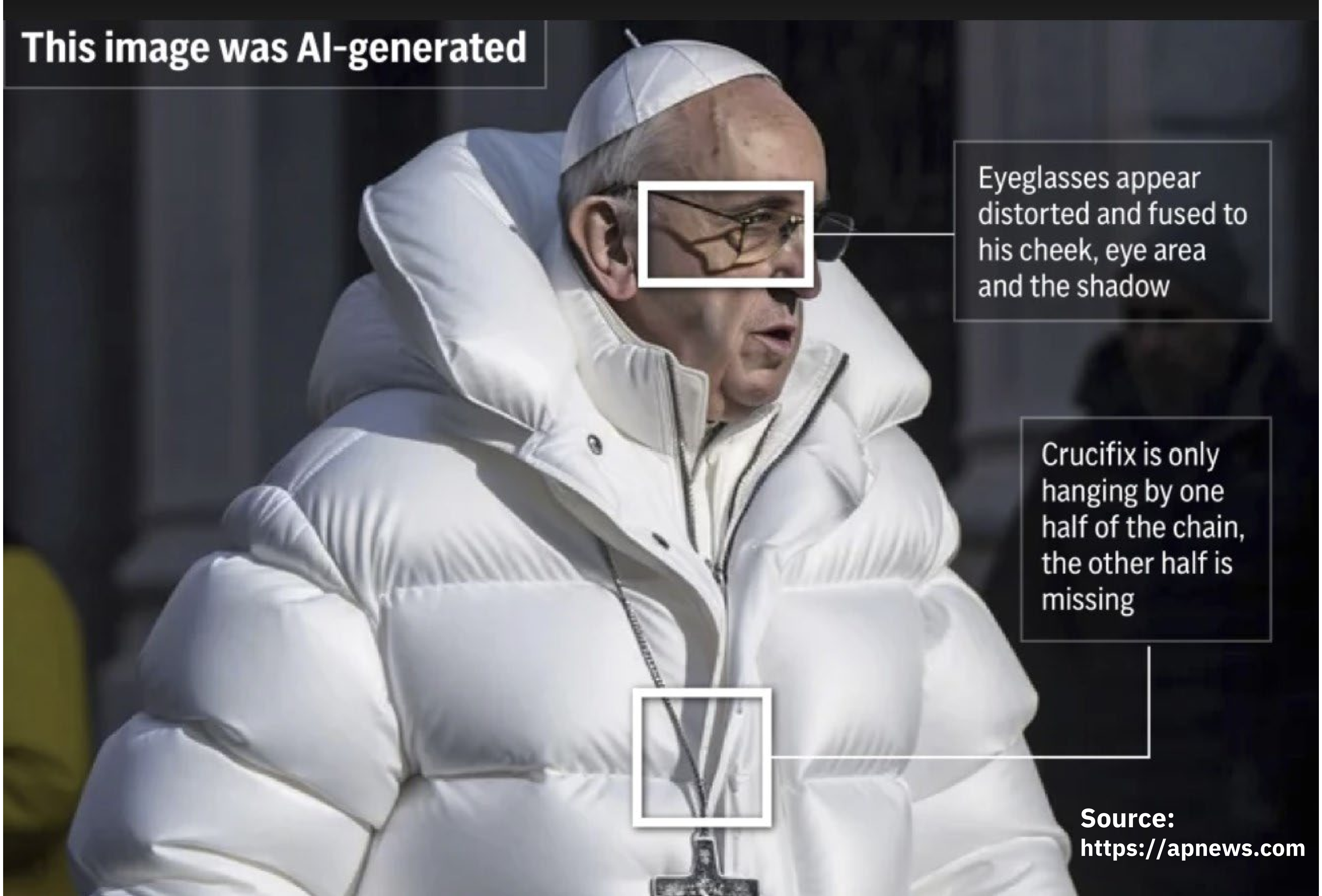
This is not Morgan Freeman – A look behind the Deepfake Singularity

https://www.youtube.com/watch?v=F4G6GNFz0O8


This is not Morgan Freeman - A look behind the Deepfake ...

YouTube · Diep Nep · Jul 29, 2021

YouTube
This is not Morgan Freeman - A look behind the Deepfake Singularity
Watch later
Share
0:02 / 1:03
YouTube



This image was AI-generated

Eyeglasses appear distorted and fused to his cheek, eye area and the shadow

Crucifix is only hanging by one half of the chain, the other half is missing

Source: https://apnews.com

Source:
https://creator.nightcafe.studio/creation/kyyQLPVhcihQ6lzlHuqu

Risks associated with Generative AI

Security

These models are susceptible to data and security risks including prompt injection attacks.

Bias

The training data has an impact on the results the model produces. Foundation Models are trained on large portions of data crawled from the internet.

Consequently, the biases that inherently exist in internet data are picked up by the trained models and can show up in the results.

Opacity

Foundation Models are also not fully auditable or transparent because of the “self-supervised” nature of the algorithm’s training.

Hallucination

LLMs can produce “hallucinations," results that satisfy a prompt syntactically but are factually incorrect.

IP

There are unanswered questions concerning the legal implications and who may own the rights to content generated by models that are trained on potentially copywritten material.

Defender’s Use of AI in Security

	Proactively Protect	**Accurately Detect**	**Accelerate Response**
Key Security Challenges	• Track high-value assets and information flow in the enterprise • Continuously evaluate risk posture to flag and remediate business impacts • Enforce policy controls to stop the intrusion, data loss, and business disruption	• Anomalous activity based on correlated telemetry from all systems & intelligence • Smarter attacks that leverage AI to evade the current generation of security controls • Increased attack surface & alert volume due to cloud and 5G	• Prioritize incidents & automate forensic activity with risk from business context • Build best practices workflows & actions based on decision-making ability • Automated interactions with the distributed environment to mitigate incidents
How can AI help?	• Enhanced data classification based on NLP techniques & deep learning • Multi-dimensional risk scoring for prioritization & continuous compliance • Automatic drift detection & policy provisioning with role mining techniques	• Continuous ML & anomaly detection (Modeling Behavioral data, Cognitive Phishing Detection) • Corroboration of threat kinetics and threat detection (Correlation) • Unsupervised ML, automated rule analysis, effective detection & recommendations	• Supervised ML based analysis for threat disposition • Graph-based analytics to investigate alerts and collaborative threat hunting • Decision support engine to compose automation rules and protection policies

Improving the productivity of security analysts with multiple foundation models

40%

of a security analyst's time is spent on automatable tasks[1]

75%

of organizations didn't identify their breaches themselves

Automate Mundane Tasks

1 **Virtual Cybersecurity Assistant**

A question-answering chatbot feature grounded on cybersecurity-specific content, providing analysts with real-time insight into an environment's specific threat landscape by asking simple questions, e.g., "Where is malicious code running in the environment?".

2 **Incident and Case Summarization**

Automating alert analysis and summarization by translating complex attack syntax to human-readable explanations of exposure, including impacted assets and recommended mitigations.

3 **Playbook Generation**

Automatically generate workflow of recommended actions for incident response and remediation using a decision support engine to compose automation rules and protection policies.

Elevate from Reactive to Proactive

4 **Threat Hunting Acceleration**

Automatically generate hunt queries from natural language that can be used to hunt for specific patterns of threats. Also use for chatbot Q&A of specific threat actors, techniques, and behaviors and cross correlate across seemingly unrelated events.

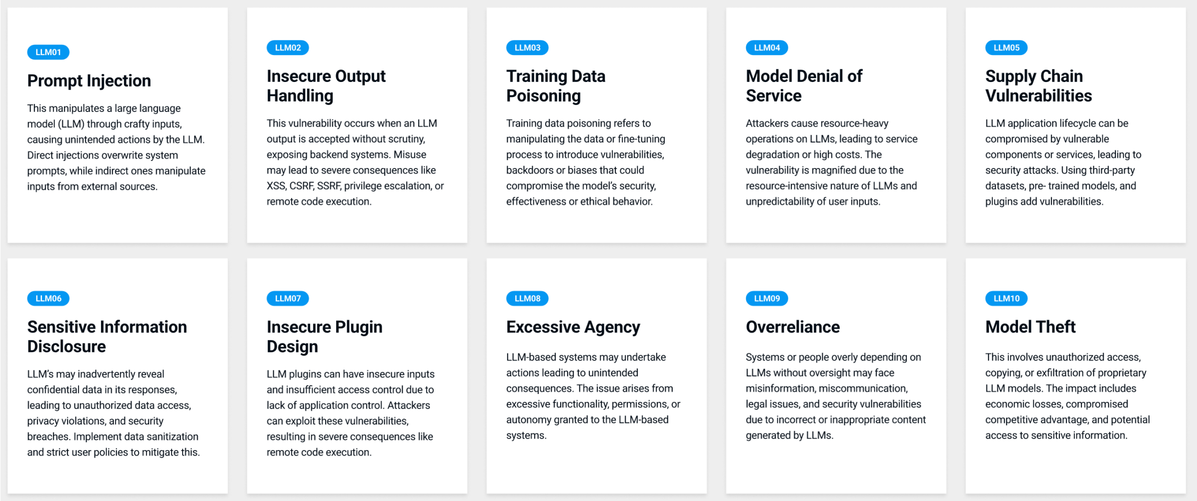
5 **Predictive Threat Insights**

Assess the possibility of a specific attack occurring, with 60-70% of attacks being "repeat offenders" (based on same code), this function will create a predictive capability that helps security teams strengthen their response readiness.

6 **Detect Previously Unseen Threats**

Identify anomalous behavior without needing to be trained on it using our continuously-learning AI foundation model that can detect and respond to previously unseen threats.

7 Best Practices to Secure AI



LLM01

Prompt Injection

This manipulates a large language model (LLM) through crafty inputs, causing unintended actions by the LLM. Direct injections overwrite system prompts, while indirect ones manipulate inputs from external sources.

LLM02

Insecure Output Handling

This vulnerability occurs when an LLM output is accepted without scrutiny, exposing backend systems. Misuse may lead to severe consequences like XSS, CSRF, SSRF, privilege escalation, or remote code execution.

LLM03

Training Data Poisoning

Training data poisoning refers to manipulating the data or fine-tuning process to introduce vulnerabilities, backdoors or biases that could compromise the model's security, effectiveness or ethical behavior.

LLM04

Model Denial of Service

Attackers cause resource-heavy operations on LLMs, leading to service degradation or high costs. The vulnerability is magnified due to the resource-intensive nature of LLMs and unpredictability of user inputs.

LLM05

Supply Chain Vulnerabilities

LLM application lifecycle can be compromised by vulnerable components or services, leading to security attacks. Using third-party datasets, pre- trained models, and plugins add vulnerabilities.

LLM06

Sensitive Information Disclosure

LLM's may inadvertently reveal confidential data in its responses, leading to unauthorized data access, privacy violations, and security breaches. Implement data sanitization and strict user policies to mitigate this.

LLM07

Insecure Plugin Design

LLM plugins can have insecure inputs and insufficient access control due to lack of application control. Attackers can exploit these vulnerabilities, resulting in severe consequences like remote code execution.

LLM08

Excessive Agency

LLM-based systems may undertake actions leading to unintended consequences. The issue arises from excessive functionality, permissions, or autonomy granted to the LLM-based systems.

LLM09

Overreliance

Systems or people overly depending on LLMs without oversight may face misinformation, miscommunication, legal issues, and security vulnerabilities due to incorrect or inappropriate content generated by LLMs.

LLM10

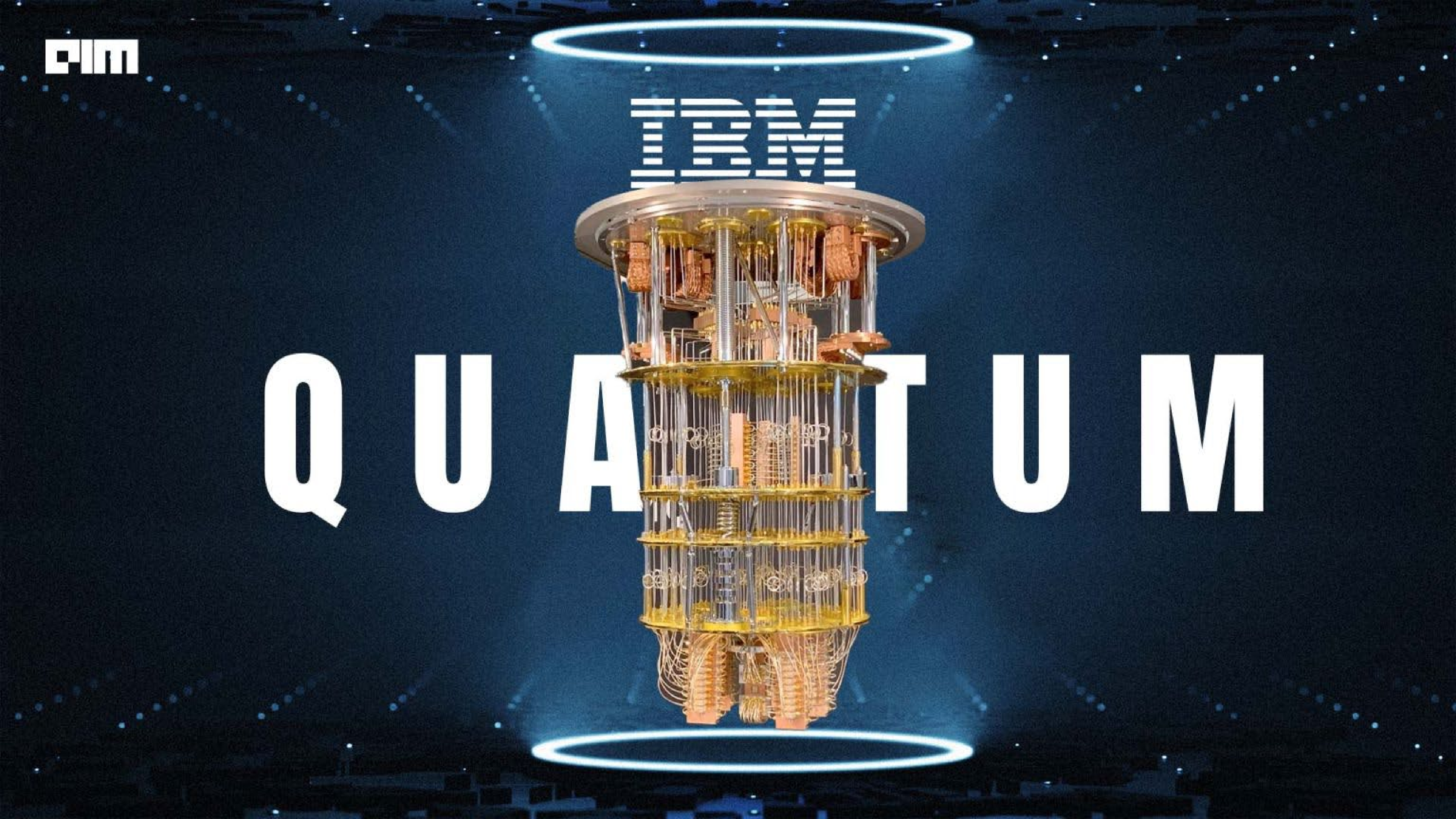
Model Theft

This involves unauthorized access, copying, or exfiltration of proprietary LLM models. The impact includes economic losses, compromised competitive advantage, and potential access to sensitive information.

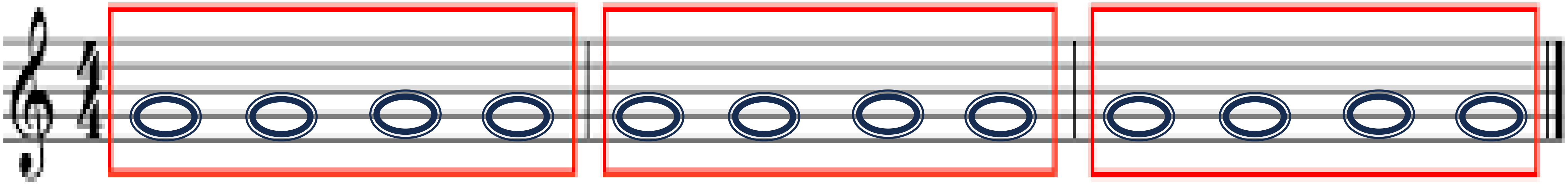
OWASP Top 10 for Large Language Model Applications

1. Leverage trusted AI by evaluating vendor policies and practices.
2. Enable secure access to users, models and data.
3. Safeguard AI models, data, and infrastructure from adversarial attacks.
4. Implement data privacy protection in the training, testing & operations phases.
5. Conduct threat modeling and secure coding practices into the AI dev lifecycle.
6. Perform threat detection & response for AI applications and infrastructure.
7. Assess and decide AI maturity through the IBM AI framework.

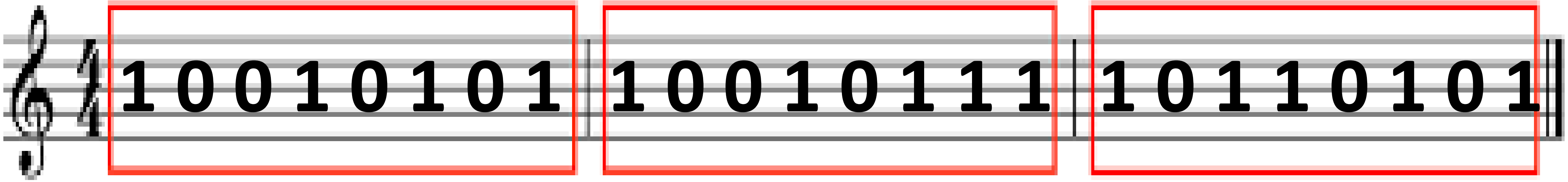
© 2023 IBM Corporation


AIM
IBM
QUANTUM

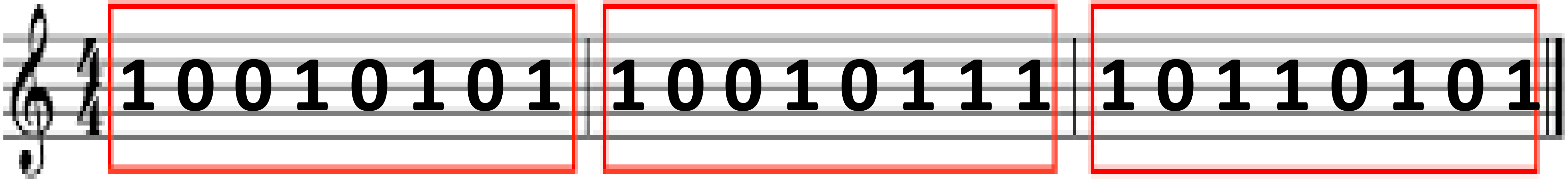
Music to Binary Bits Concept

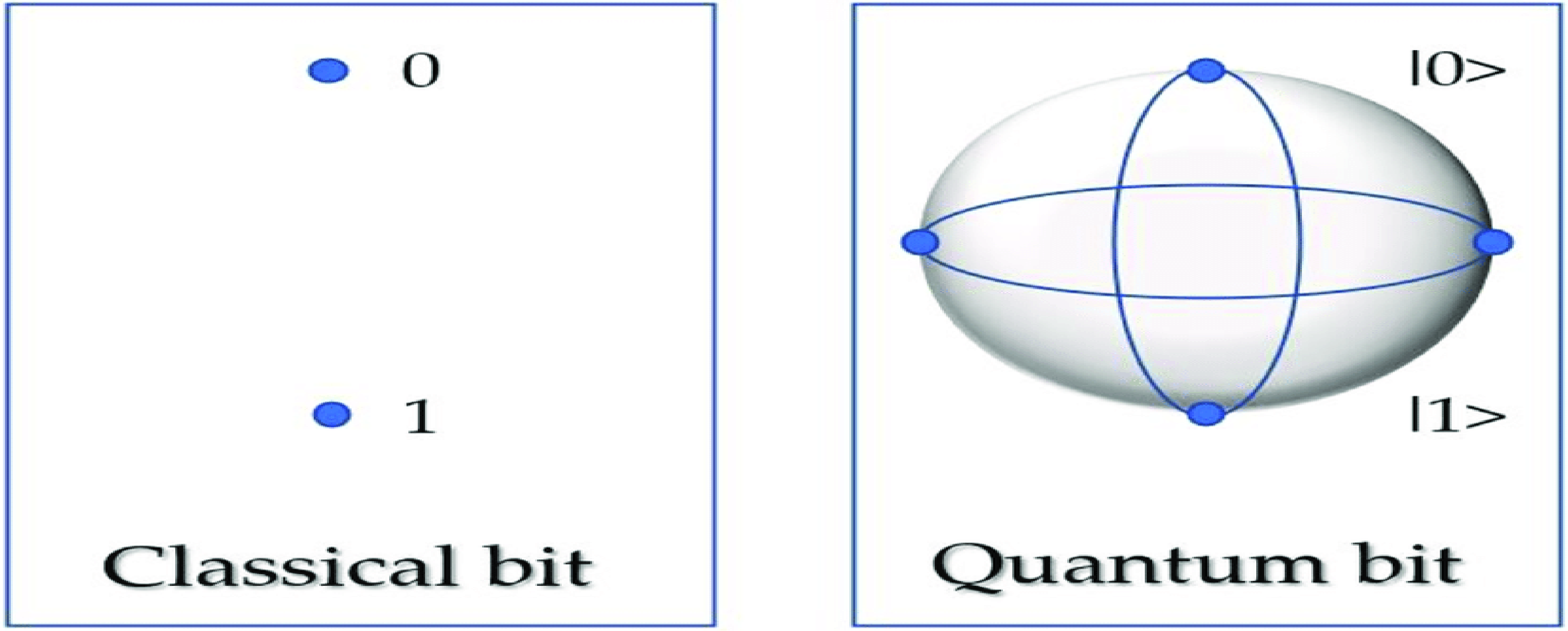


Music to Binary Bits Concept



Music to Binary Bits Concept

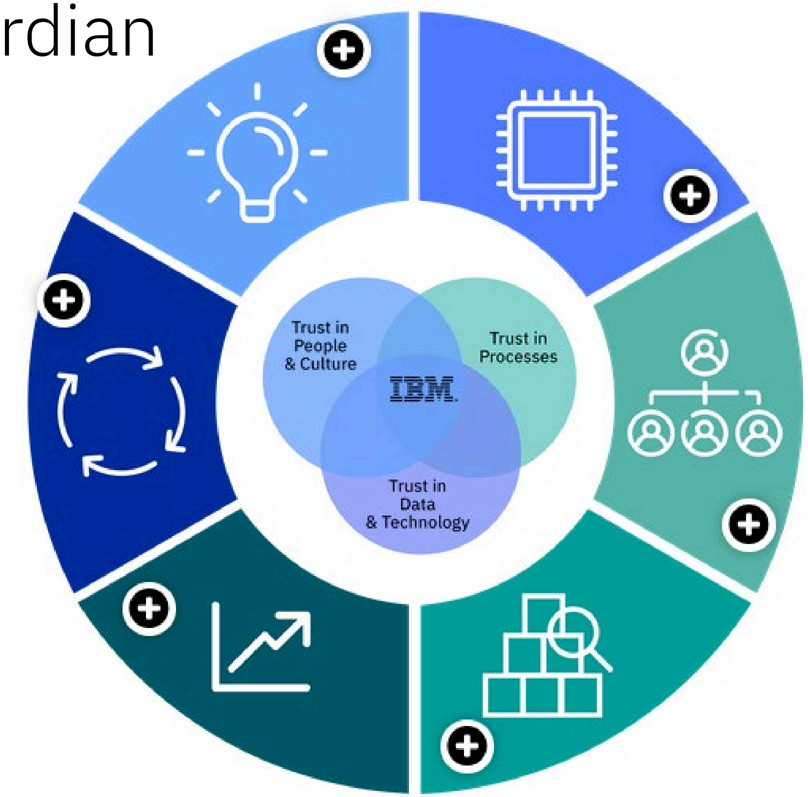




https://www.researchgate.net/figure/Classical-bit-vs-Quantum-bit-Q-bit_fig1_348386747

Never Trust, Always Verify

Make your public information public
Keep your private information private
Treat AI as an entity with access control
Assign a guardian



As a Process

- Conduct risk assessment (initially and regularly) and mitigate.
- Consider current and upcoming data and AI regulations.
- Review models, practices, processes, structures, skills, and culture for trustworthiness.
- Support and automate internal audits on input and output.

ML + Sec + OPS

Cost of a Data Breach 2024 Report

$9.36M

US average cost of a data breach (global $4.88M)

82%
Breaches were based on Cloud data

$4.99M
Average cost of a malicious insider attack

$1.4M

Complex Security Environments

$2.2M

Cost savings – from extensive AI in prevention automation

1 in 3

Breaches involving shadow data

$1M
Cost Savings when Law Enforcement is involved in ransomware attacks

Top Mitigating Factors

- Discover and protect data across cloud environments
- Security AI and Automation
- Adopt an attacker’s perspective of your organization’s environment

Time to identify and contain

US average: 216 days ($2.25M savings under 200 Days)

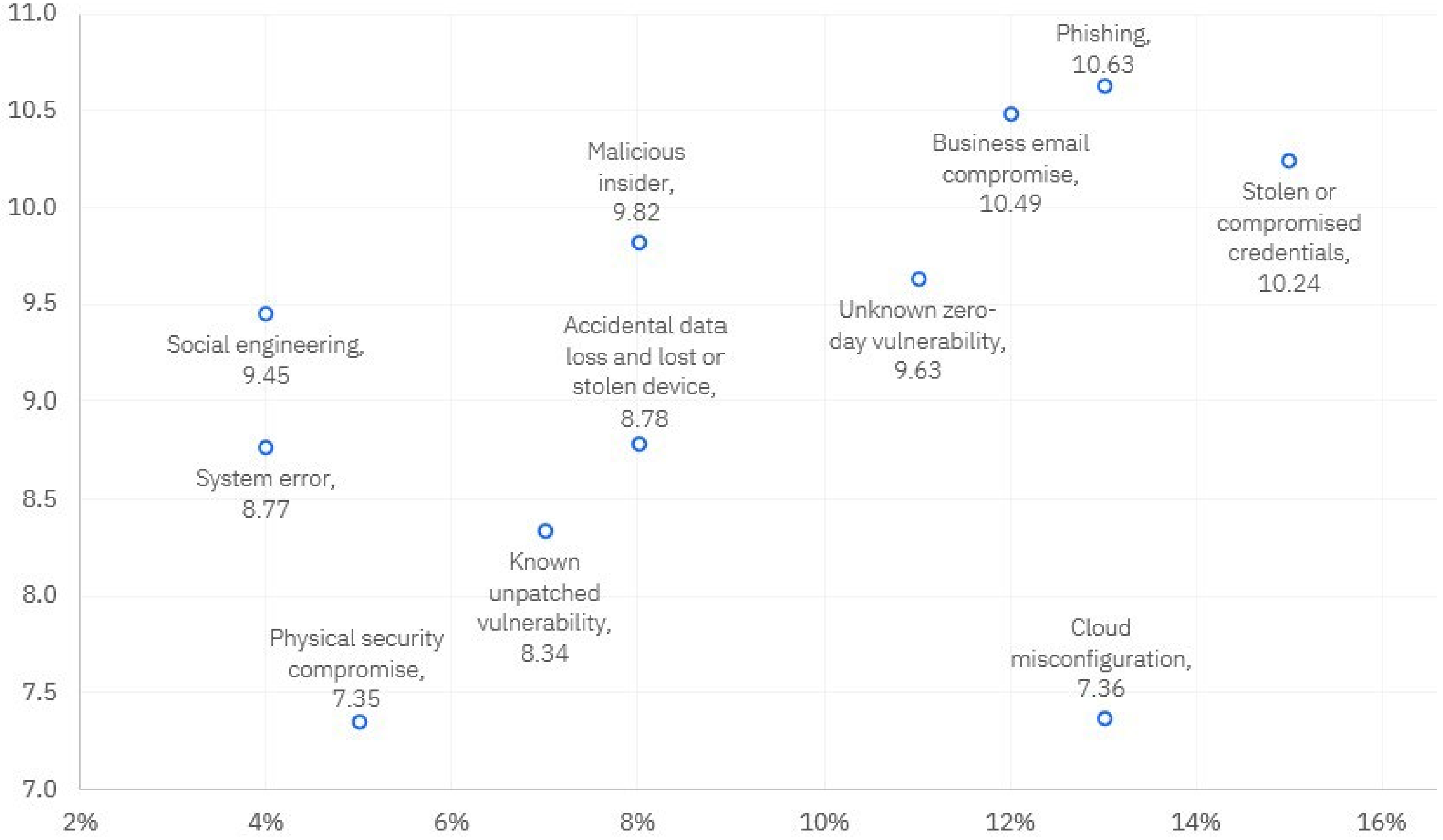
169 Identify

47 Contain

$1.49 million ⬇
Average cost savings with incident response teams and IR testing vs. low levels in IR teams & testing

Total cost and frequency of data breaches by initial attack vector

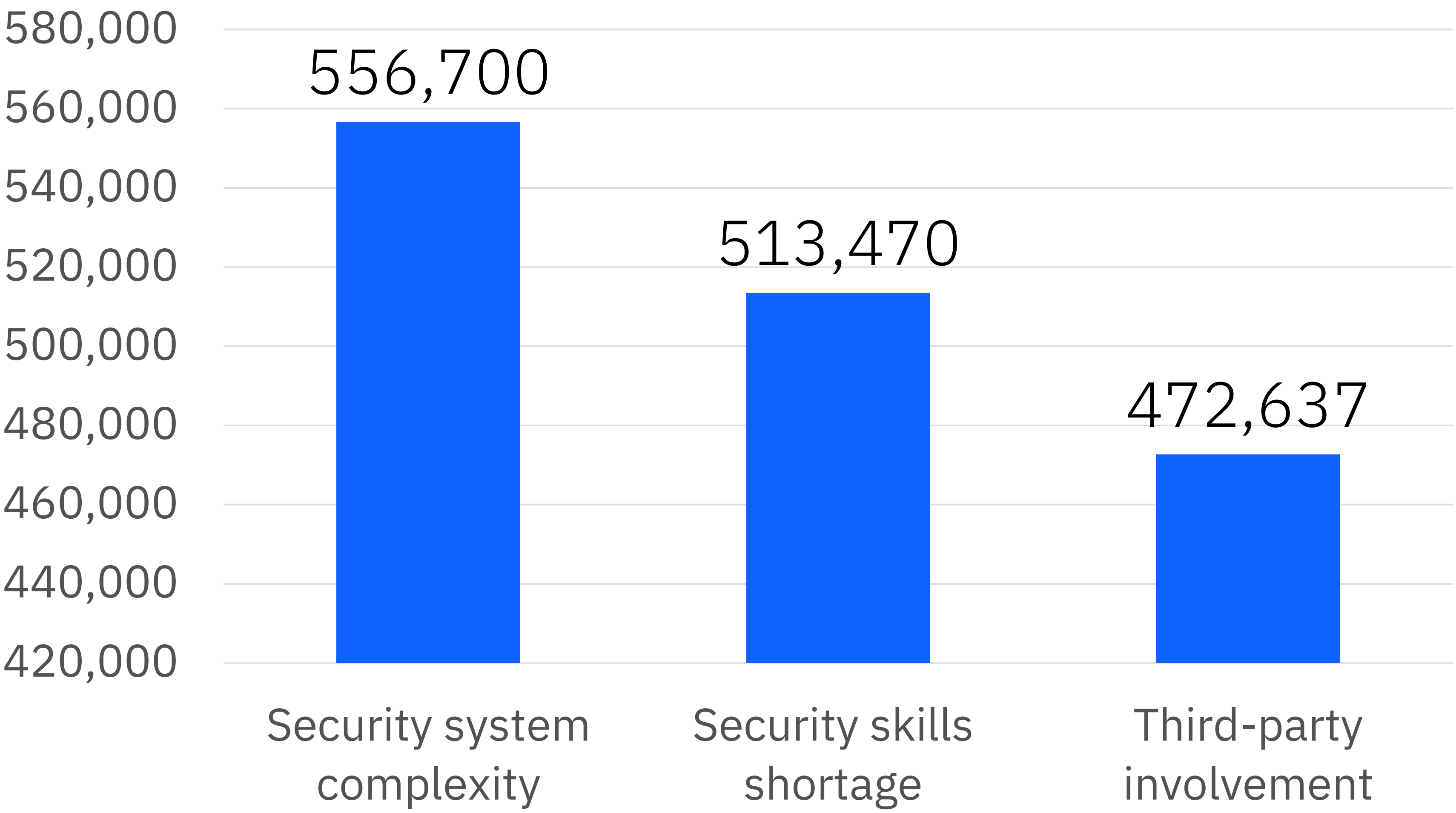
Measured in USD millions



© 2024 IBM Corporation

Factors that may increase the cost of a US data breach

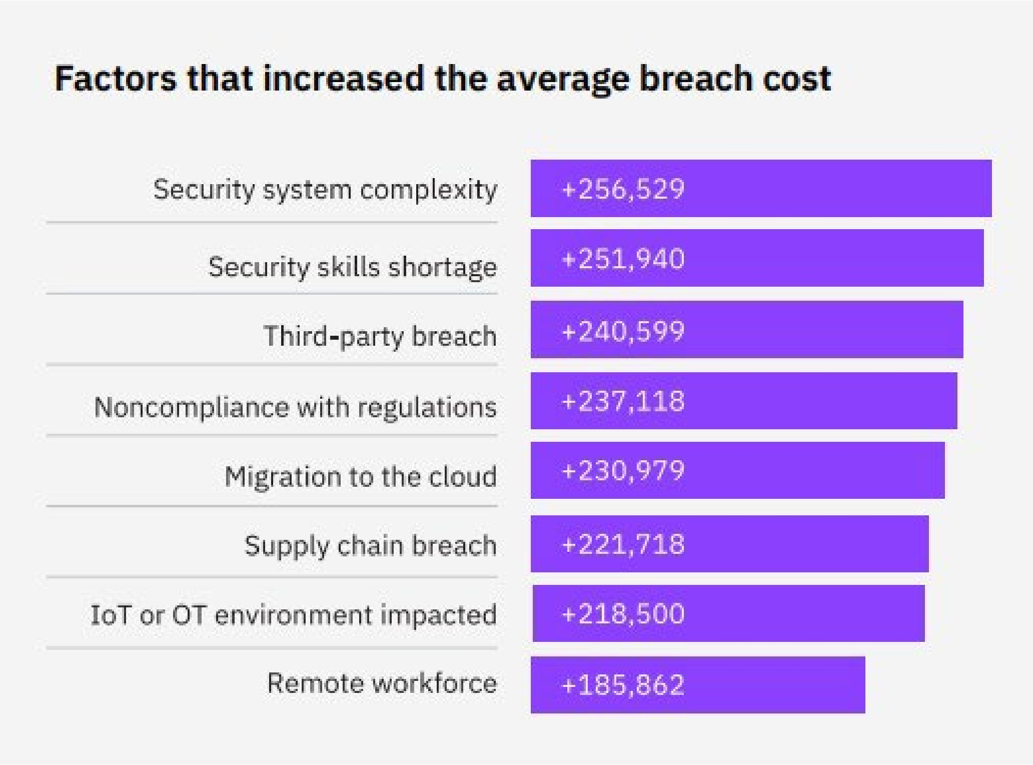
Measured in USD

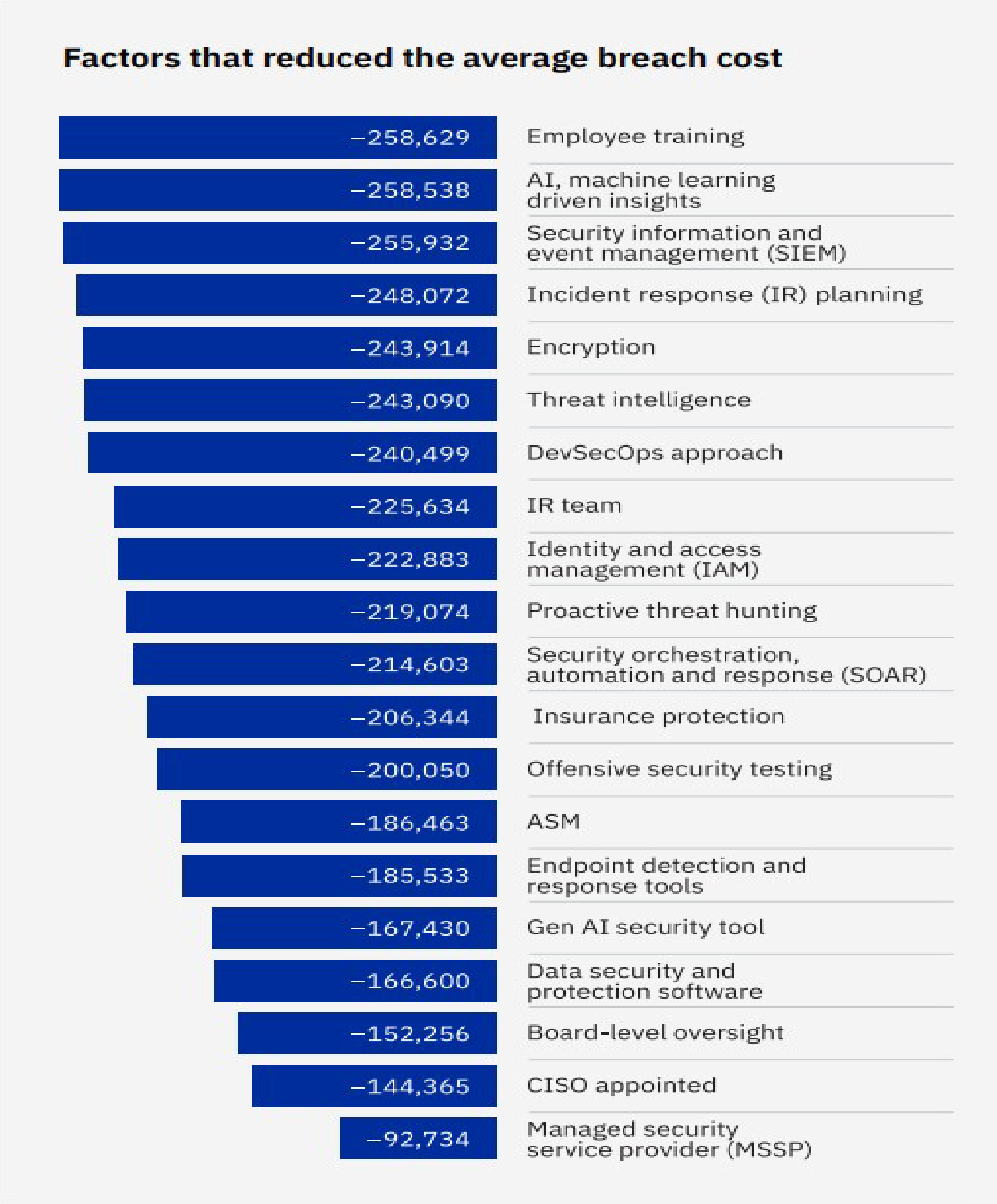


© 2024 IBM Corporation

Factors that Increase and Decrease the cost of a data breach

Measured in USD


Factors that increased the average breach cost


Factors that reduced the average breach cost

© 2024 IBM Corporation

The reality of expanding attack surfaces.

30%

of assets are unknown or unmanaged due to rapid transformation. [1]

76%

Organizations have been compromised by an unknown or unmanaged asset. [2]

50%

By 2026, non-patchable attack surfaces will grow to account for more than half of an enterprise’s total exposure. [3]

---

[1] Forrester Attack Surface Management

[2] 2023 ESG Security Hygiene and Posture Management Remains Decentralized and Complex

[3] Gartner, Implementing A Continuous Threat Exposure Management Program

Help Socialize: Lockdown Your Credit Reporting

- https://www.equifax.com/
- https://www.experian.com/
- https://www.transunion.com/
- https://Innovis.com/

https://www.linkedin.com/in/melore/
@MichaelMelore
mmelore@us.ibm.com
www.SecRT.us