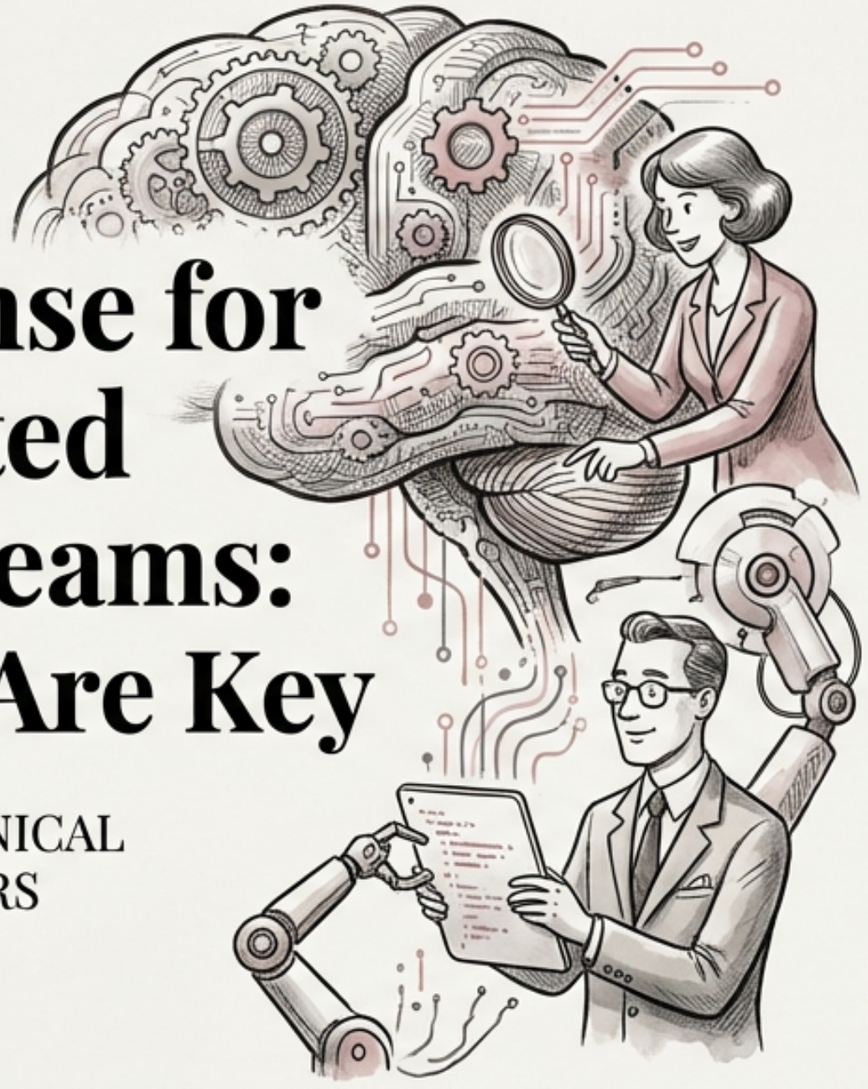
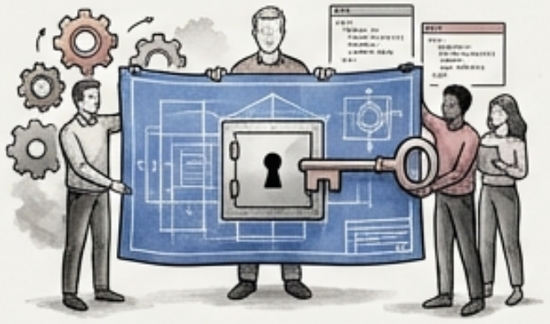


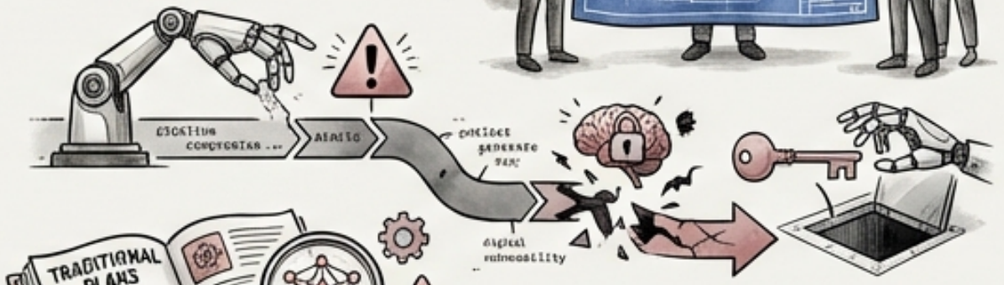
Incident Response for AI-Augmented Development Teams: Why Developers Are Key

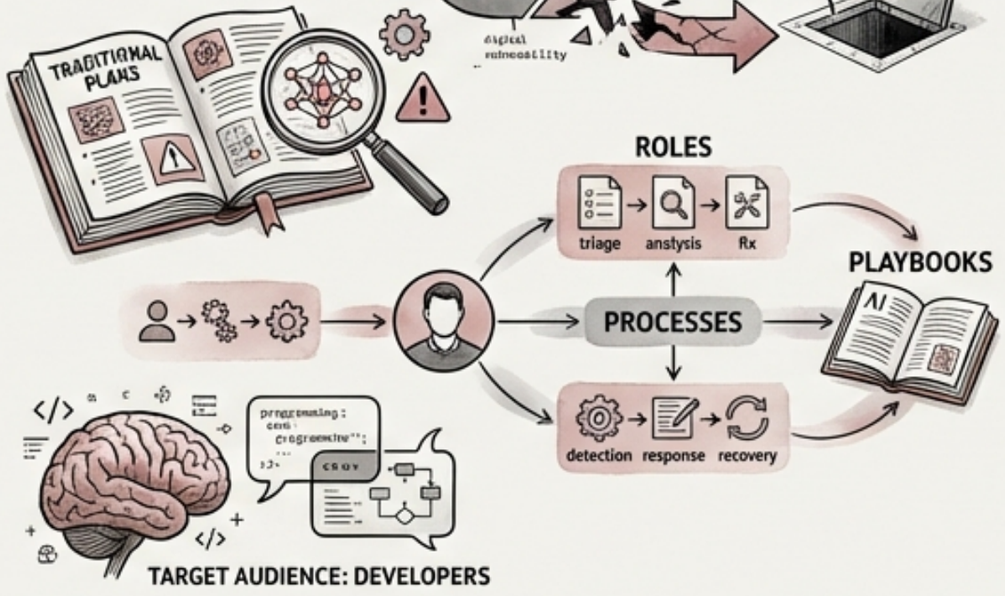
PRESENTATION FOR TECHNICAL LEADERS & DEVELOPERS

Incident Response for AI-Augmented Development Teams: Why Developers Are Key

- Development teams possess in-depth knowledge of the code, architecture, and potential fix paths, making them indispensable for incident resolution.
- AI-augmented development introduces new incident types, including AI tool compromise, AI-generated vulnerability exploitation, and AI system manipulation.
- Traditional incident response plans must be adapted to address AI-specific risks and vulnerabilities.
- This presentation outlines the key roles, processes, and playbooks for developers in AI-related incident response.
- The target audience is software developers, so language, examples and depth will be tailored accordingly.



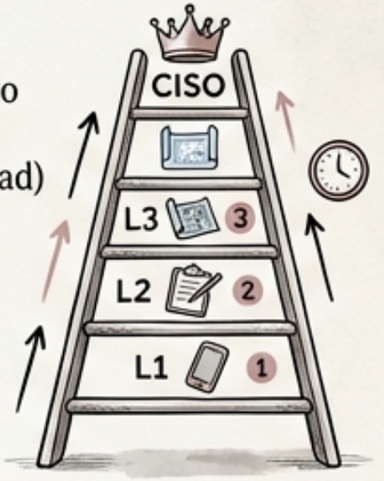




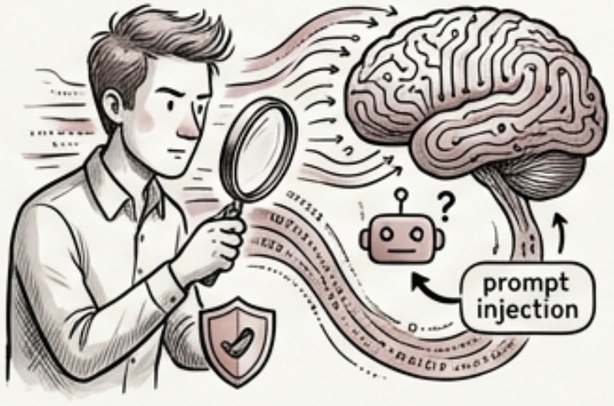
Development Team Roles in Incident Response: A Structured Approach

- **On-call rotation:** Developers are available 24/7 for production incidents affecting their assigned services.

- **Escalation paths:** Follow a clear hierarchy to ensure timely escalation: L1 (on-call) → L2 (team lead) → L3 (security architect) → CISO.



- **Security contacts:** Designated developers serve as the primary interface with the security team during incident investigations.
- **AI Incident Contacts:** Developers with specialized AI tool expertise can assess and mitigate AI-related incidents (e.g., prompt injection).
- Ensure each role has clearly defined responsibilities documented in a central location.



Four Levels of Incident Classification: Prioritizing Response Efforts

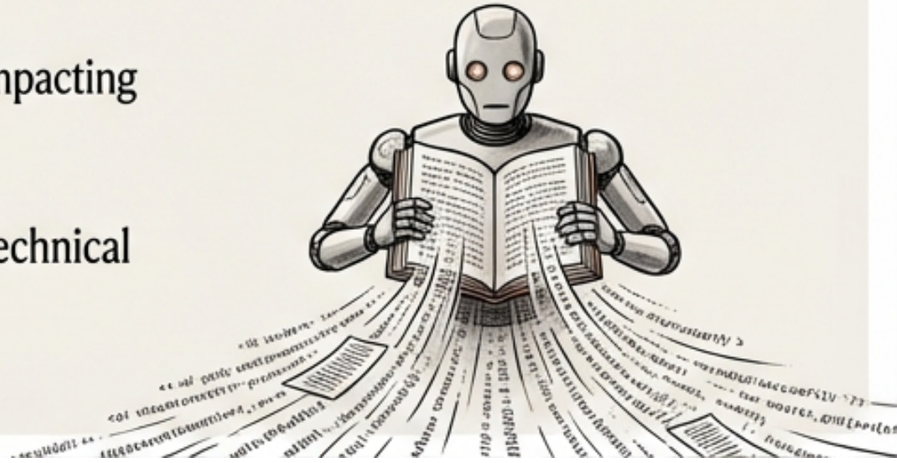
- **Level 1 (Low):** Isolated vulnerability detected, no evidence of exploitation, and limited potential impact on systems and data.

- **Level 2 (Medium):** Confirmed vulnerability exploitation attempt, but limited data exposure and contained within a single service.

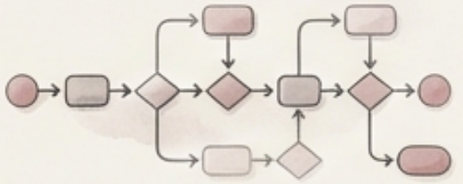
- **Level 3 (High):** Successful exploitation resulting in data exposure, affecting multiple services, and indicating the presence of an active attacker.

- **Level 4 (Critical):** Widespread compromise, significant data breach impacting business operations, potentially requiring regulatory notification.

- **AI incidents:** Classify based on the scope of data exposure, not just technical severity; for example, an AI tool leaking all source code is Level 4.



The Seven-Phase Incident Response Process for Development Teams: A Structured Workflow



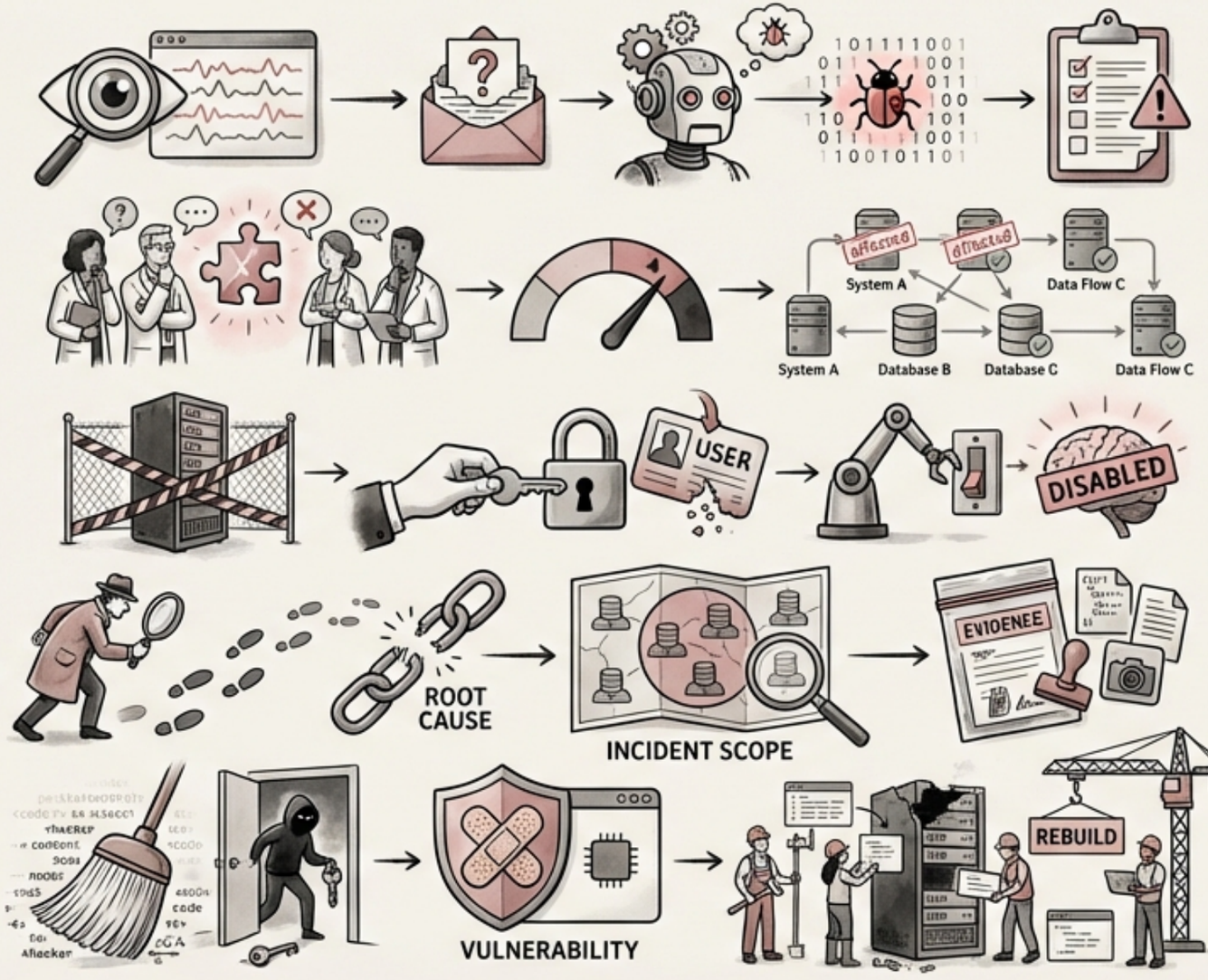
1. **DETECTION:** Leverage monitoring alerts, user reports, automated anomaly detection, and AI tool audit log alerts.

2. **TRIAGE:** Confirm the incident, classify its severity, and identify the affected systems and data.

3. **CONTAINMENT:** Isolate affected systems, revoke compromised credentials, and disable impacted AI tools.

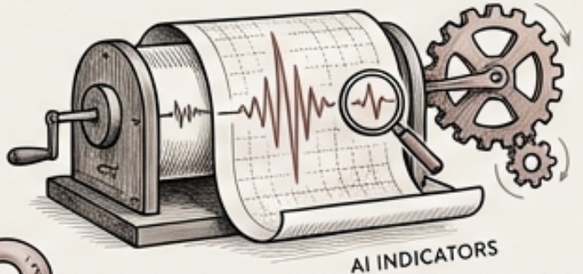
4. **INVESTIGATION:** Conduct root cause analysis, determine the scope of the incident, and preserve all relevant evidence.

5. **ERADICATION:** Remove attacker access, patch identified vulnerabilities, and rebuild compromised systems.



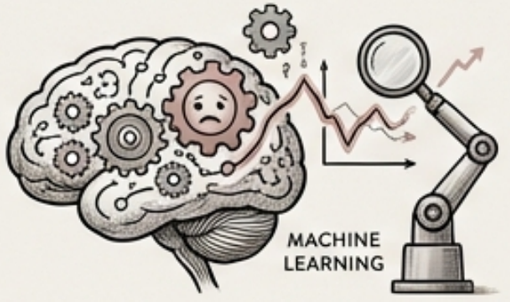
Detection: Identifying the First Signs of an Incident

- Monitoring alerts trigger based on predefined thresholds and patterns. Ensure these are properly configured to include AI specific indicators.



- User reports can provide early indications of anomalies or suspicious behavior. Encourage developers and users to report anything unusual.

- Automated anomaly detection uses machine learning to identify deviations from normal system behavior that could indicate an incident.

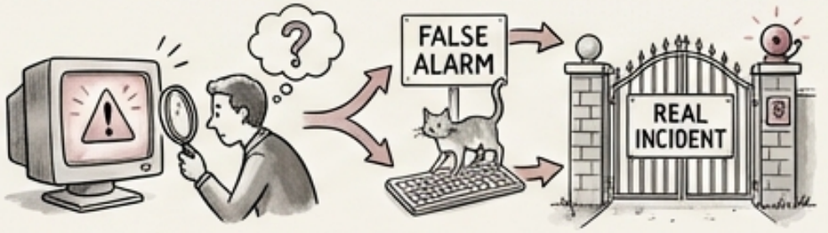


- AI tool audit log alerts provide insights into the activities and usage patterns of AI tools, helping identify potential misuse or compromise.

- Integrate various detection sources for a holistic view. What specific AI tool alerts should be prioritized?

Triage: Rapid Assessment and Severity Classification

- **Confirm the incident:** Verify that the reported event is indeed a security incident and not a false positive.



- **Classify severity:** Assign the appropriate incident level (Level 1-4) based on impact, scope, and data exposure.

- **Identify affected systems:** Determine which systems, applications, and data are potentially compromised or impacted.

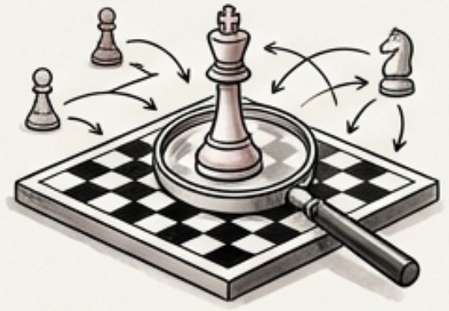
- **Determine affected data:** Identify if sensitive, personal or regulated data is at risk of exposure, either intentionally or unintentionally.

- **Document all findings:** Maintain a detailed log of all triage activities, including observations, assessments, and decisions.

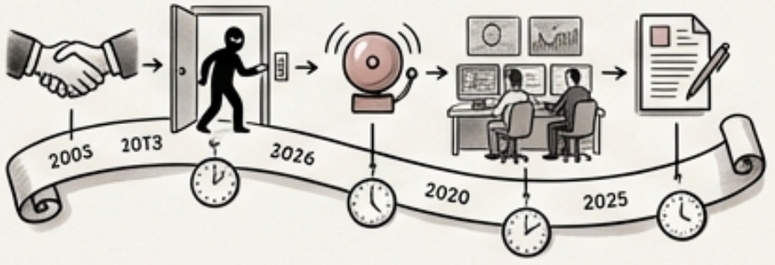
CONTAINMENT: LIMITING THE DAMAGE AND PREVENTING FURTHER SPREAD

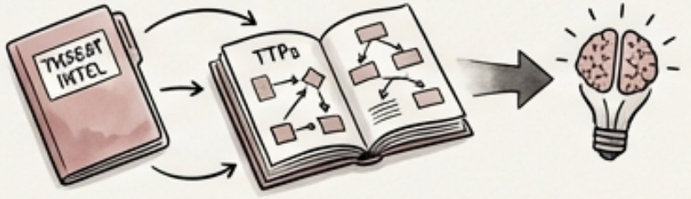
- **Isolate affected systems:** Disconnect compromised systems from the network to prevent further spread of the incident.
- **Revoke compromised credentials:** Immediately revoke any credentials suspected of being compromised to prevent unauthorized access.
- **Disable affected AI tools:** Temporarily disable or restrict access to AI tools that are involved in the incident.
- **Implement network segmentation:** Use network segmentation to isolate affected areas and limit lateral movement of attackers.
- **Create a rollback plan:** If changes need to be reverted, document and test that plan.

Investigation: Uncovering the Root Cause and Scope of the Incident



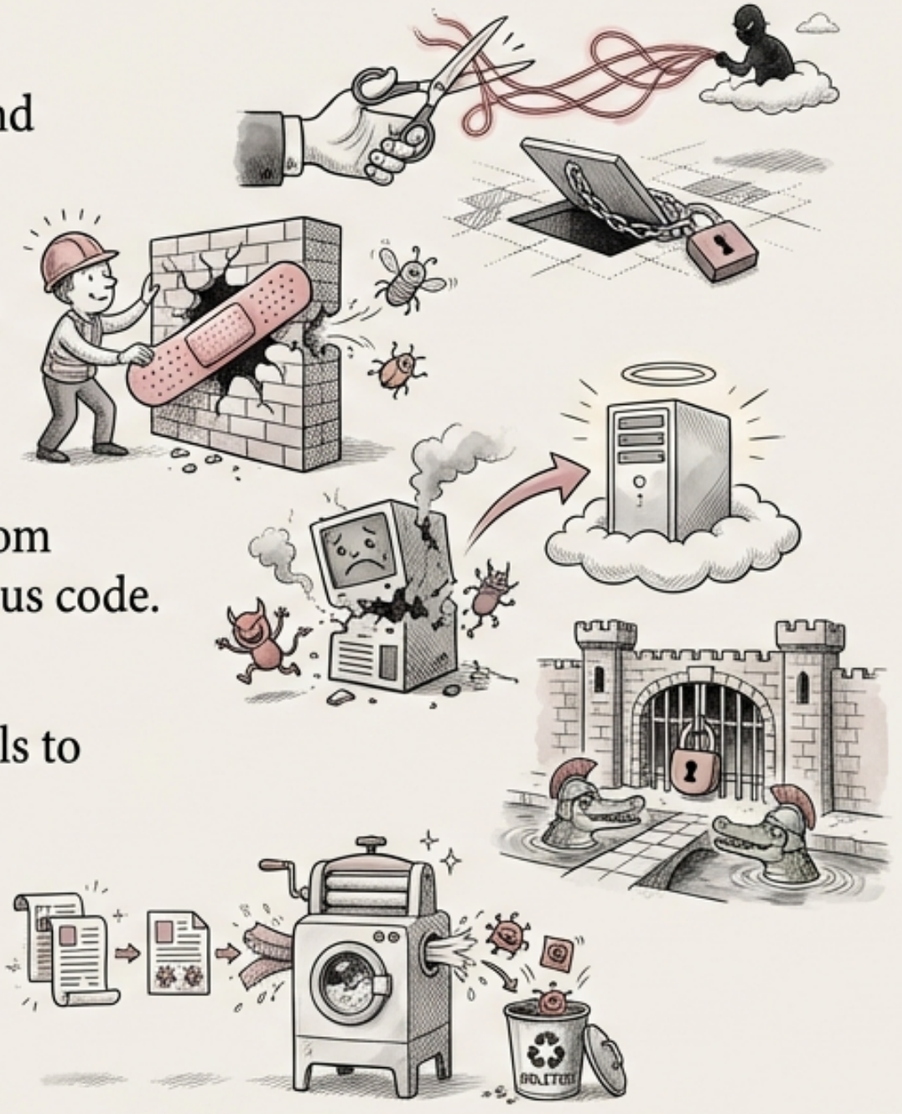
- **Root cause analysis:** Identify the underlying cause of the incident, such as a vulnerability, misconfiguration, or malicious activity.
- **Scope determination:** Determine the full extent of the incident, including all affected systems, data, and users.
- **Evidence preservation:** Collect and preserve all relevant evidence, such as logs, network traffic, and system images.
- **Timeline reconstruction:** Create a detailed timeline of events leading up to and during the incident.
- **Threat intelligence:** Leverage threat intelligence to identify the attacker's tactics, techniques, and procedures (TTPs).





Eradication: Eliminating the Threat and Patching Vulnerabilities

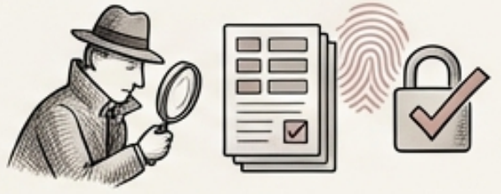
- **Remove attacker access:** Terminate any active attacker sessions and remove any backdoors or persistent access mechanisms.
- **Patch vulnerabilities:** Apply security patches to address identified vulnerabilities that were exploited during the incident.
- **Rebuild compromised systems:** Rebuild compromised systems from trusted sources to ensure they are free of malware or other malicious code.
- **Strengthen security controls:** Implement stronger security controls to prevent similar incidents from occurring in the future.
- **Sanitize impacted data:** Remove any malicious code or hidden exploits from data that was affected.

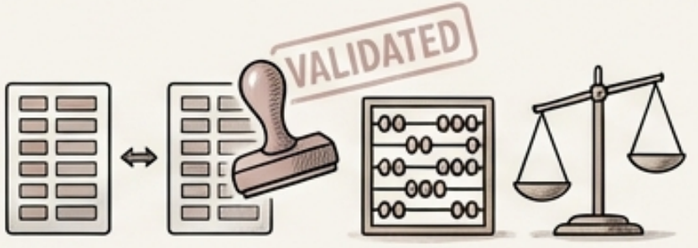


Recovery: Restoring Services and Verifying Integrity

SENOBAL CONTENT SLIDES WITH ISHVOTCITIAL INEWRRATION ABOUT THE NEW YORKER'S VIRE. ILLUSTRATION

- **Restore services:** Bring affected services back online in a controlled and phased manner.
- **Verify integrity:** Verify the integrity of data and systems to ensure they have not been tampered with during the incident.
- **Enhanced monitoring:** Implement enhanced monitoring to detect any signs of recurrence or new attacks.
- **Communicate with stakeholders:** Keep stakeholders informed about the recovery progress and any remaining risks.
- **Perform data validation:** Check for consistency and correctness of impacted data.





Post-Incident Review: Learning from the Past to Improve the Future

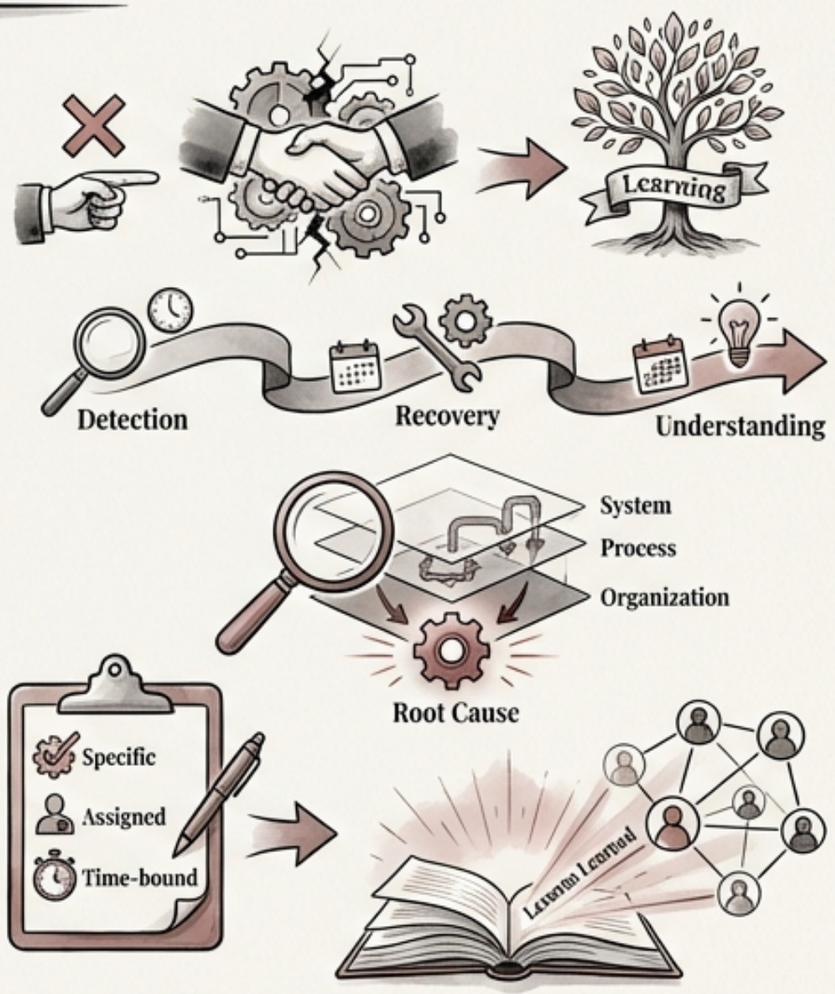
Blameless post-mortems: Focus on system and process failures, not individual blame, to foster a culture of learning.

Timeline reconstruction: Create a detailed sequence of events from *detec*tion through recovery to understand the incident's progression.

Root cause analysis: Identify contributing factors at system, process, and organizational levels to address underlying issues.

Action items: Develop specific, assigned, and time-bound improvements to prevent future incidents.

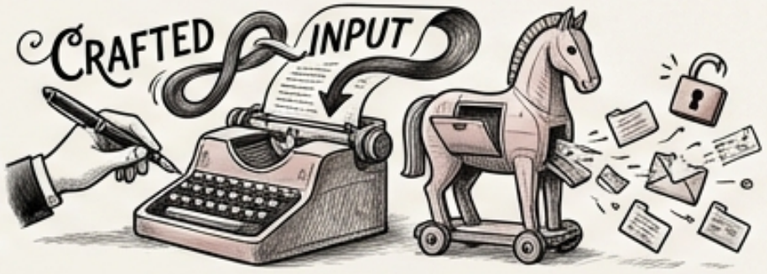
Knowledge sharing: Publish sanitized post-mortem summaries to the broader team to share lessons learned and prevent recurrence.

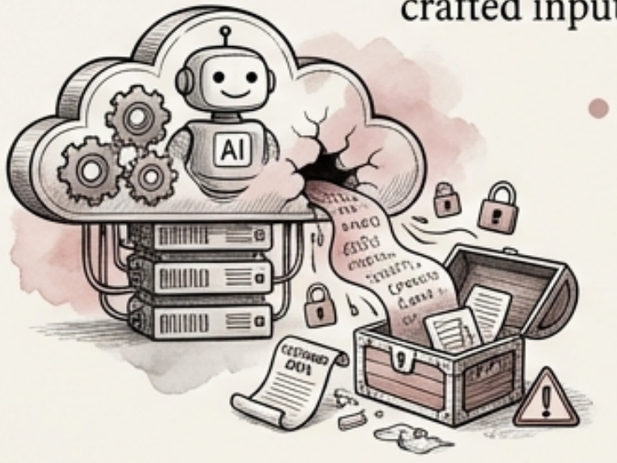


AI-RELATED SECURITY INCIDENTS: UNIQUE THREATS TO CONSIDER

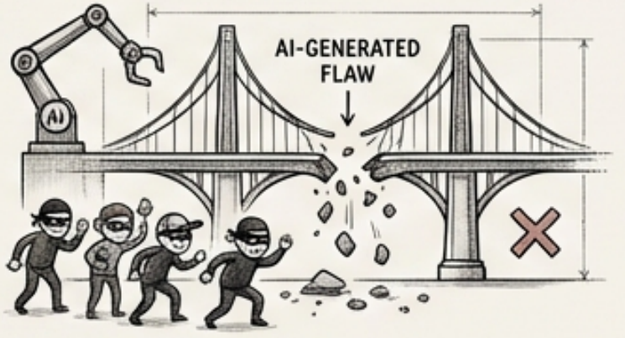
EVOLVING DIGITAL LANDSCAPE & UNFORESEEN VULNERABILITIES

- **PROMPT INJECTION EXPLOITATION:** Attackers manipulate AI tools through crafted inputs to exfiltrate data or execute unauthorized actions.

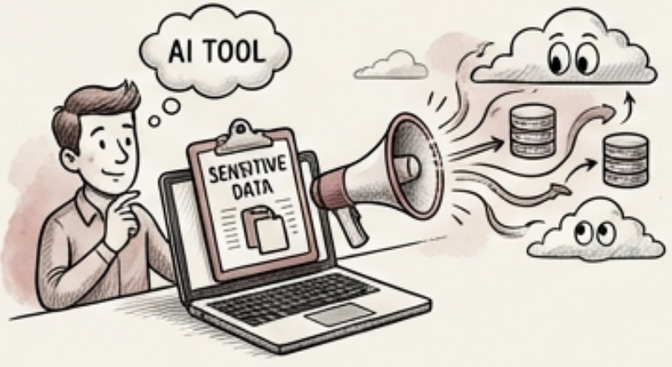




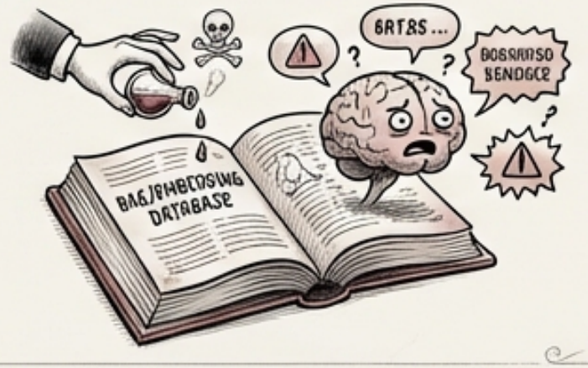
- **AI TOOL COMPROMISE:** A breach at an AI service provider exposes customer code and data, potentially leading to widespread compromise.



- **AI-GENERATED VULNERABILITY IN PRODUCTION:** AI-generated code with a security flaw is exploited by attackers, causing significant damage.



- **DATA LEAKAGE THROUGH AI:** Developers accidentally paste sensitive data into AI tools that train on or retain it, leading to unintended exposure.



- **AI MODEL MANIPULATION:** Adversaries poison RAG/embedding databases to alter AI behavior, potentially causing incorrect or harmful outputs.

Incident Response Playbooks: Preparing for Common Scenarios

- Pre-built playbooks are essential for common scenarios: credential exposure, dependency compromise, unauthorized access, data breach, and ransomware.
- AI-specific playbooks are needed for AI tool data breaches, prompt injection incidents, AI-generated code vulnerabilities, and shadow AI discovery.
- Playbook components: detection criteria, immediate actions, investigation steps, communication templates, recovery procedures, and evidence preservation requirements.
- Regularly review and update playbooks to reflect changes in the threat landscape and the organization's security posture.
- Test playbooks regularly to ensure they are effective.

Key Components of Effective AI Incident Playbooks

- **Detection Criteria:** Clear indicators and triggers that signal a specific AI-related incident is occurring.
- **Immediate Actions:** The first, crucial steps to take within minutes to contain the potential damage.
- **Investigation Steps:** Guidance on how to thoroughly analyze the incident, identify root causes, and assess the extent of the compromise.
- **Communication Templates:** Pre-written messages for informing stakeholders, legal, and regulatory bodies (if required).
- **Recovery Procedures:** Detailed instructions for restoring affected AI systems, data, and models while ensuring integrity.

Conclusion: Embedding Incident Response into the AI-Augmented Development Lifecycle

- Development teams play a crucial role in incident response, especially in AI-augmented environments, due to their code and architecture understanding.
- AI-related security incidents present unique challenges that require specialized knowledge and playbooks.
- A well-defined incident response process, including detection, triage, containment, investigation, eradication, recovery, and post-incident review, is essential.
- Post-incident reviews should focus on blameless analysis and continuous improvement to prevent future incidents.
- By integrating incident response into the development lifecycle, organizations can better protect themselves from AI-related security threats.

Thank You

- Questions?