RESEARCH PROJECT PROPOSAL

SECURITY EVENT LOG ANONYMIZATION AND SYNTHETIC ATTACK GENERATION FOR REALISTIC SECURITY ANALYST TRAINING AND INTRUSION DETECTION RESEARCH

OVERVIEW

- Project Requirements
- Research Questions & Contribution
- Aims & Objectives
- Literature Review
- Research Design & Artefacts
- Ethics & Risk
- Project Timeline

PROJECT REQUIREMENTS

- Mandatory deliverable for MSc Cyber Security (Essex, n.d.)
- British Computer Society (BCS) major projects criteria (BCS, 2022):
  - Demonstrate practical work using computing/IT technology
  - Problem definition and research objectives
  - Final report including results, critical appraisal and lessons learned
- Cyber Security Body of Knowledge (CYBOK) knowledge area topic (Essex, n.d)

REQUIREMENTS – DELIVERABLES TRACEABILITY

Requirement	Project activity or artefact
Demonstrate advanced understanding of problem and existing research	Literature Review Research Question Development
Demonstrate originality in knowledge application	Pursuit of anonymization over AI for data generation
Demonstrate technical skill	Design and program an ETL pipeline
Demonstrate critical thinking and communication skills	Identity and articulate deanonymization risk likelihood
Demonstrate consideration of legal and ethical matters	Data privacy requirement reviews
CYBOK Knowledge Area based research	8.3 – Analysis methods 5.1 – Privacy as confidentiality 6.4 – Malware detection

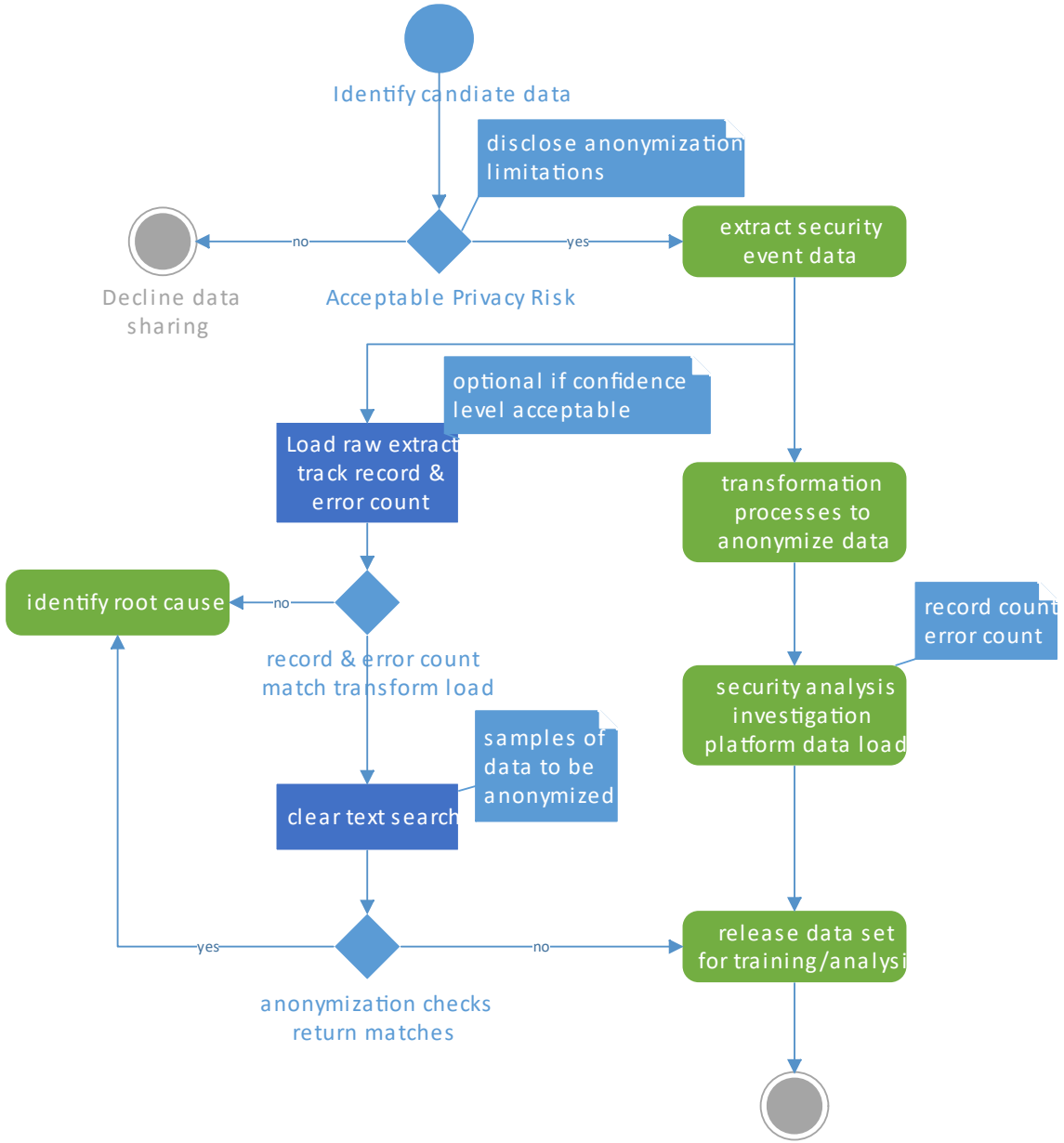
CONTRIBUTION

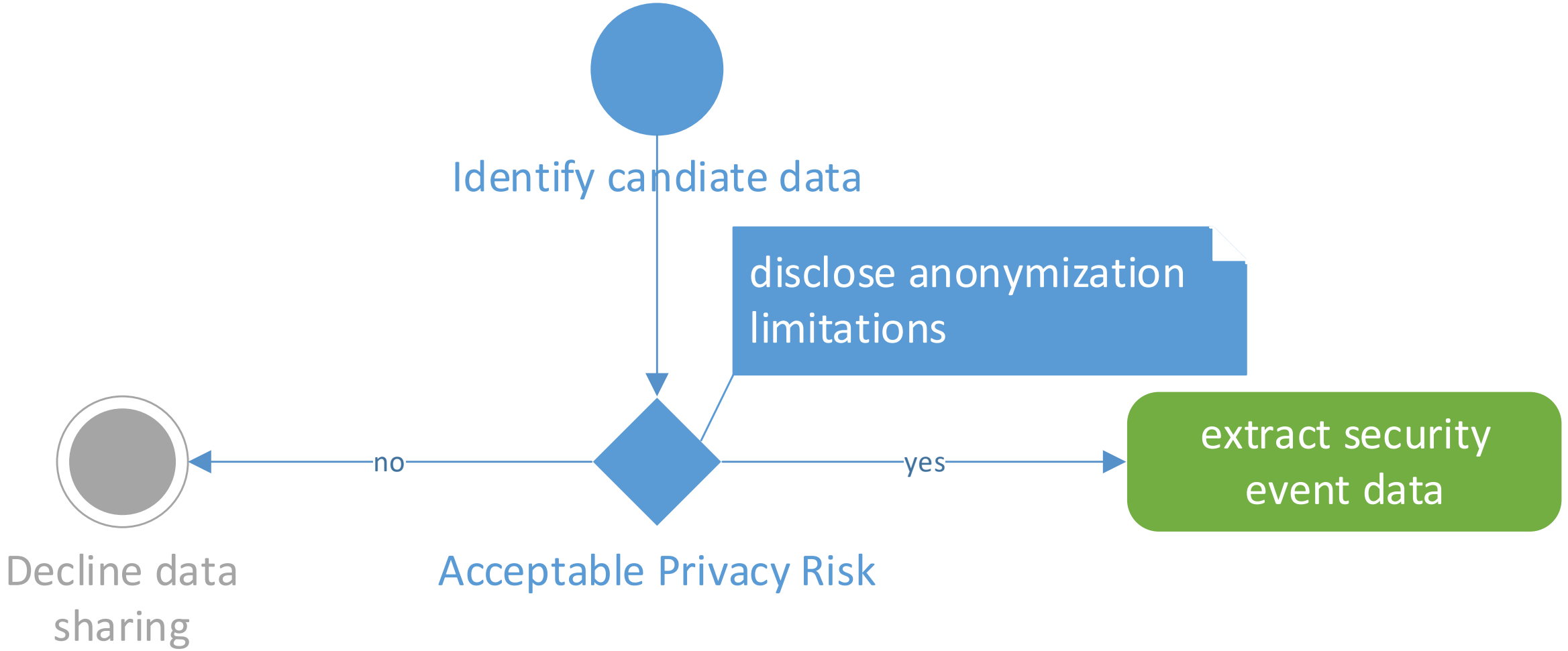
- *"…specifics of the problem domain…(iv) the enormous variability of benign traffic, making it difficult to find stable notions of normality…" (Sommer & Paxson, 2010)*
- *"…training sets must be created to enable cybersecurity research so that new AI tools can be developed to enhance the effectiveness of current cybersecurity analysts…" (Bresniker et al, 2019)*

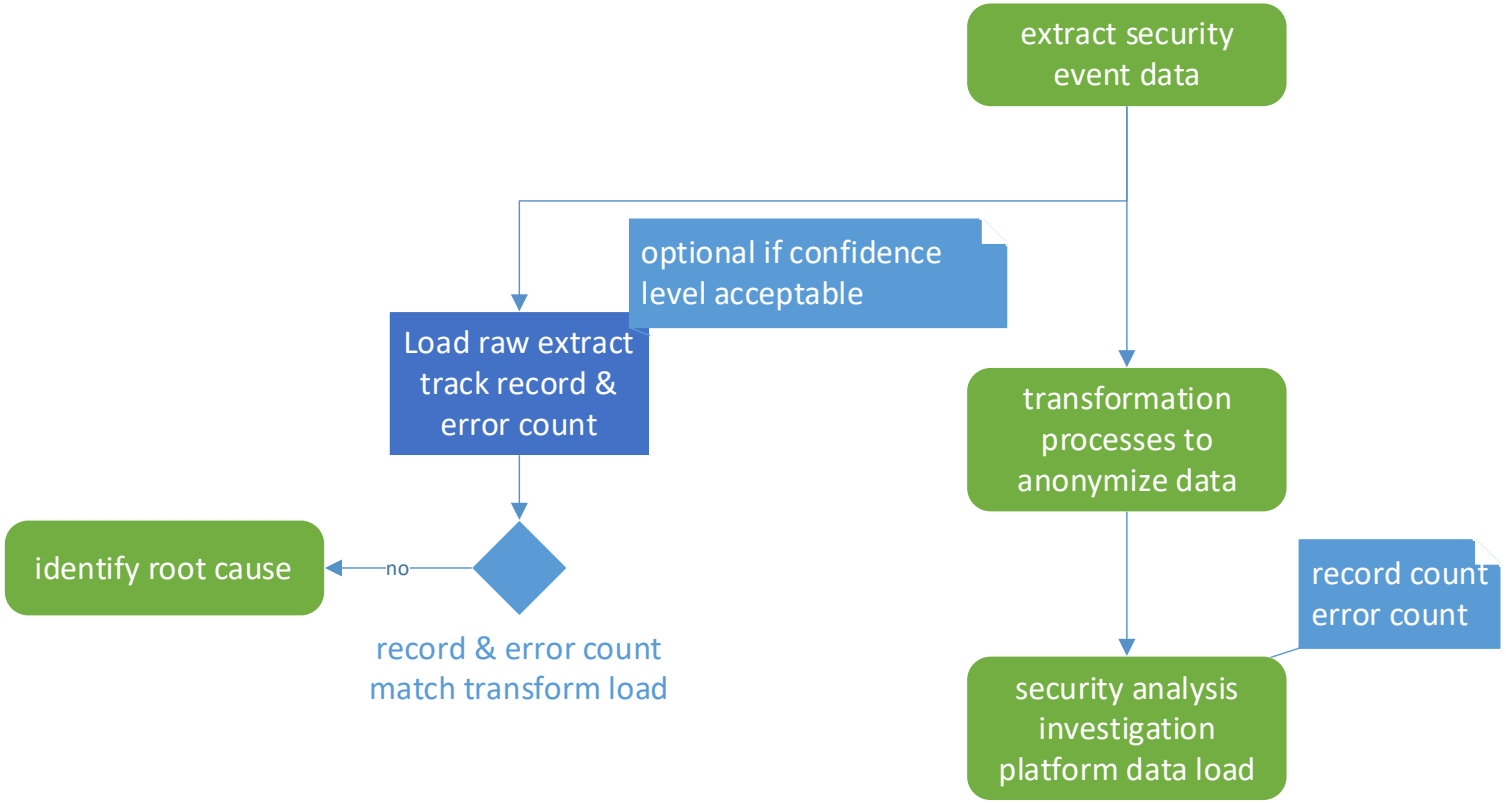
- *The proposed project supports cyber security training and research by providing:*
  - *a mechanism for enabling access to current realistic security event data*
  - *An approach to maintaining confidentiality for the original data owner*
  - *A mechanism to inject new attack indicators into an anonymized data set*

RESEARCH QUESTIONS

- How can security event logs be programmatically anonymized while retaining the key characteristics and data volumes needed for cyber security analyst intrusion detection skills training and research?
- What log features must be anonymized to prevent identification of people or critical systems?
- What protection measures must be applied to anonymized, highly repeatable, and structured data to reasonably prevent reversing attacks?
- How can synthetic log events for newly documented attacks be generated as realistic input into an existing security event monitoring platform?

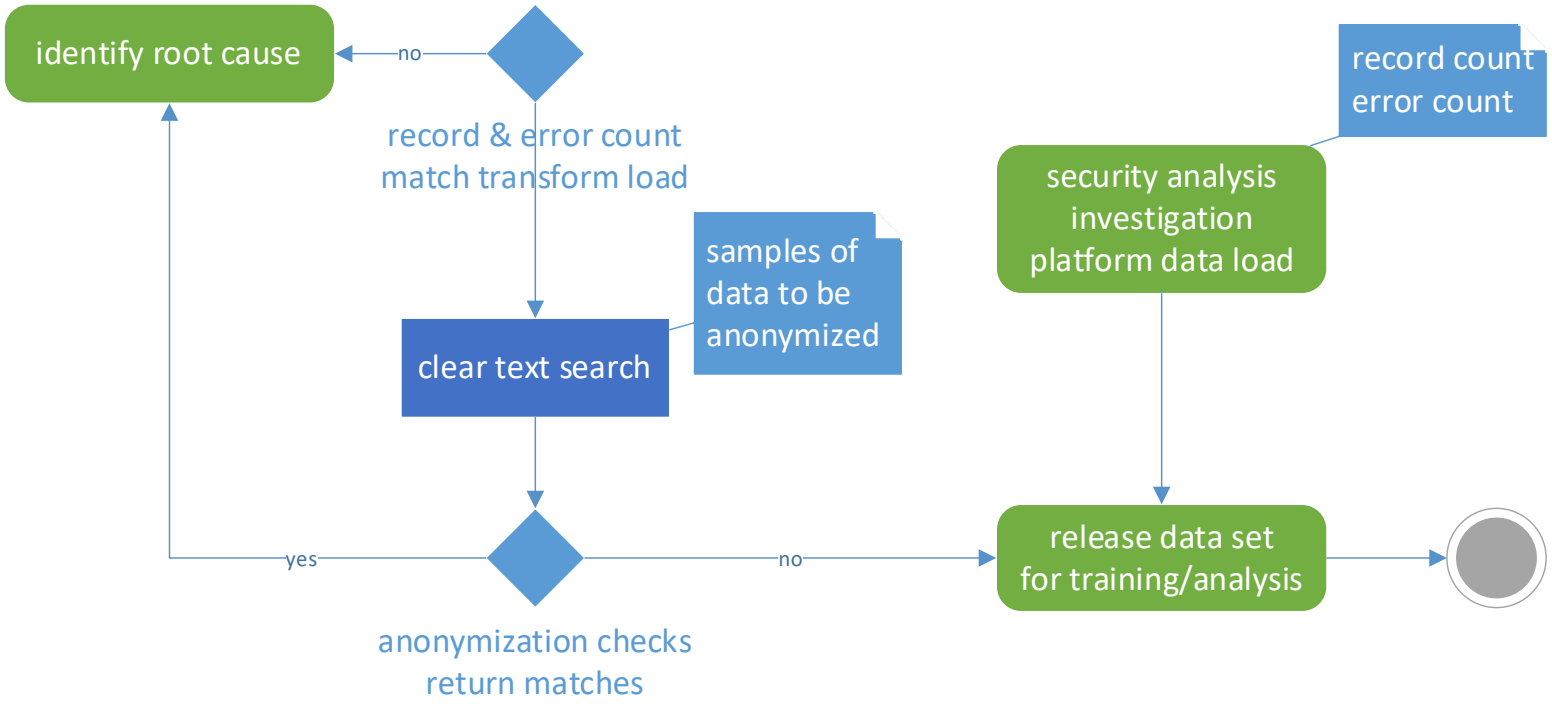

Identify candiate data
disclose anonymization limitations
no
yes
Decline data sharing
Acceptable Privacy Risk
extract security event data
optional if confidence level acceptable
Load raw extract track record & error count
transformation processes to anonymize data
identify root cause
no
record & error count match transform load
record count error count
security analysis investigation platform data load
samples of data to be anonymized
clear text search
yes
no
release data set for training/analysi
anonymization checks return matches


Identify candiate data
disclose anonymization limitations
no
yes
extract security event data
Decline data sharing
Acceptable Privacy Risk



Aims & Objectives: ETL pipeline

- Data owner controls the extraction process
- Transformation process software run by either party
- Load process can support initial testing:
  - Same record count
  - Same error on load count
- Validation can be optional once confidence achieved
- Output format must match input for maximum utilization

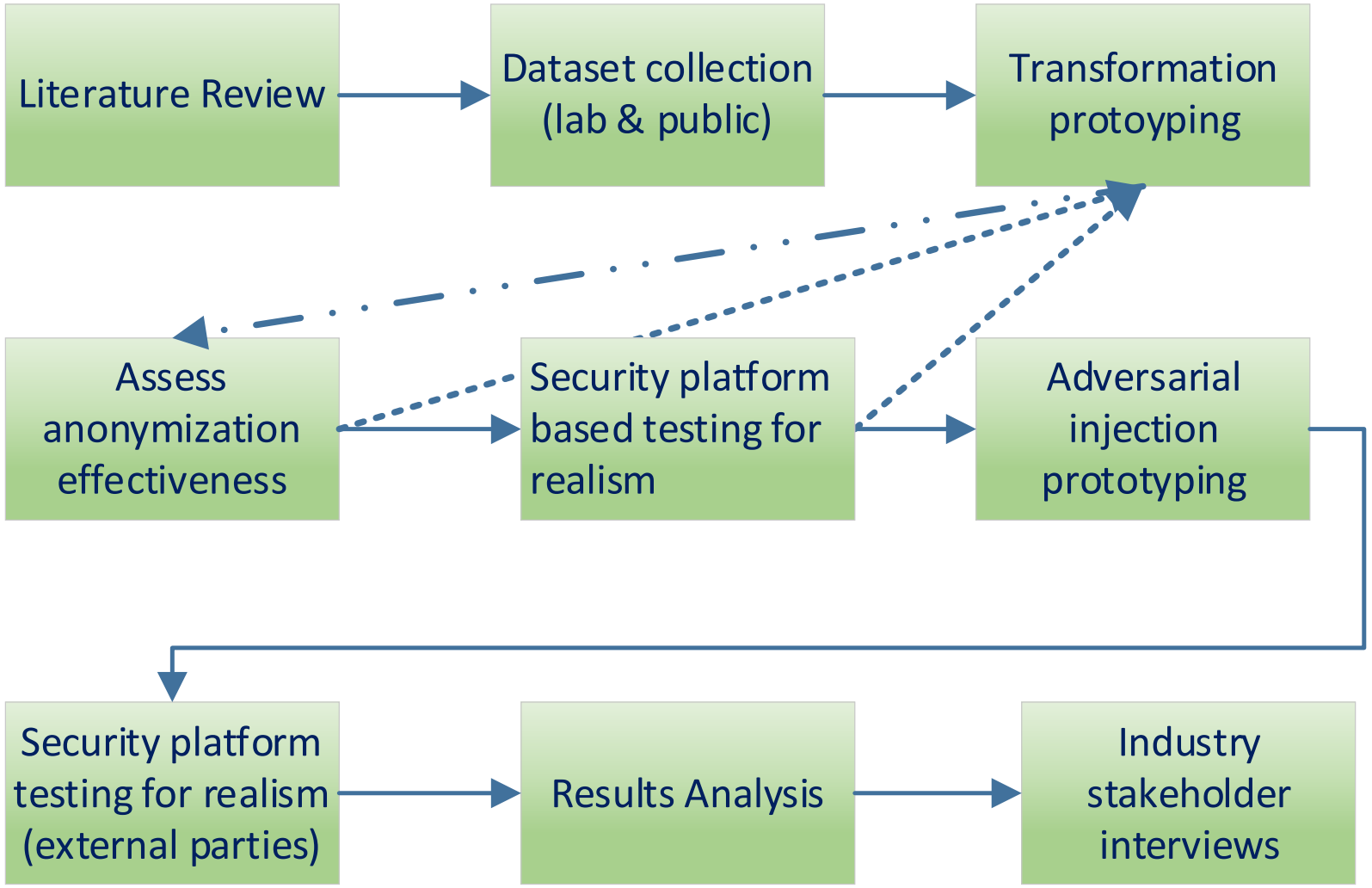


Aims & Objectives: ETL ctn'd

- Load process can support initial testing:
  - Same record count
  - Same error on load count
- Secondary anonymization checks:
  - Search capabilities within platform
  - Full record searches raw string matching
  - Byte comparison

LITERATURE REVIEW

Key words & searching	Relevance
Synthetic log generation Generative adversarial network (GAN) Artificial Intelligence, AI, ML, DL	- Identify limitations and potential algorithms or frameworks that can be used for testing - Identify other publicly know datasets that may be suitable for ETL pipeline software development
Security event log syntax / labeling	- Identify which log elements are strong candidates for anonymization as labeling can differ between sources
Anomoly / intrusion detection	- The primary function of a cyber security analyst or supporting toolset - Identify leading research and commercial options
Adversarial simulation	- Identify key features of frameworks and approaches for modeling cyber attacks to ensure correct data generation
Data anonymization / deanonymization	- Both a critical success factor and constraint



Project Design Overview

ETHICAL CONSIDERATIONS

- "*To share or not to share, that is the dilemma …"*
  - Potential inference/deanonymization attacks
  - Log content legally collected as part of security measures
  - Limited personally identifiable information
  - High level of diversity E.G., Apache http access and error logs, Windows Active Directory events

- Shortage of skilled cyber security professionals (ISC2, 2021)
- Realistic training may improve analysts' ability to perform in real environments

RISK MANAGEMENT

- Some public datasets include both benign and documented attack events in raw formats suitable for developing initial transformation programs and testing for attack events known to exist
- Hardware and software needed to develop a cyber range and analysis laboratory is in place, no external financial support is required
- Multiple years of professional experience working with security event logs and investigation platforms and holds multiple SANS GIAC certifications related to this research area
- Many information security community connections from which to draw testing participants
- Extra time allocated to riskiest stages of the project

PROJECT TIMELINE

Security Event Log Anonymization

Essex final project

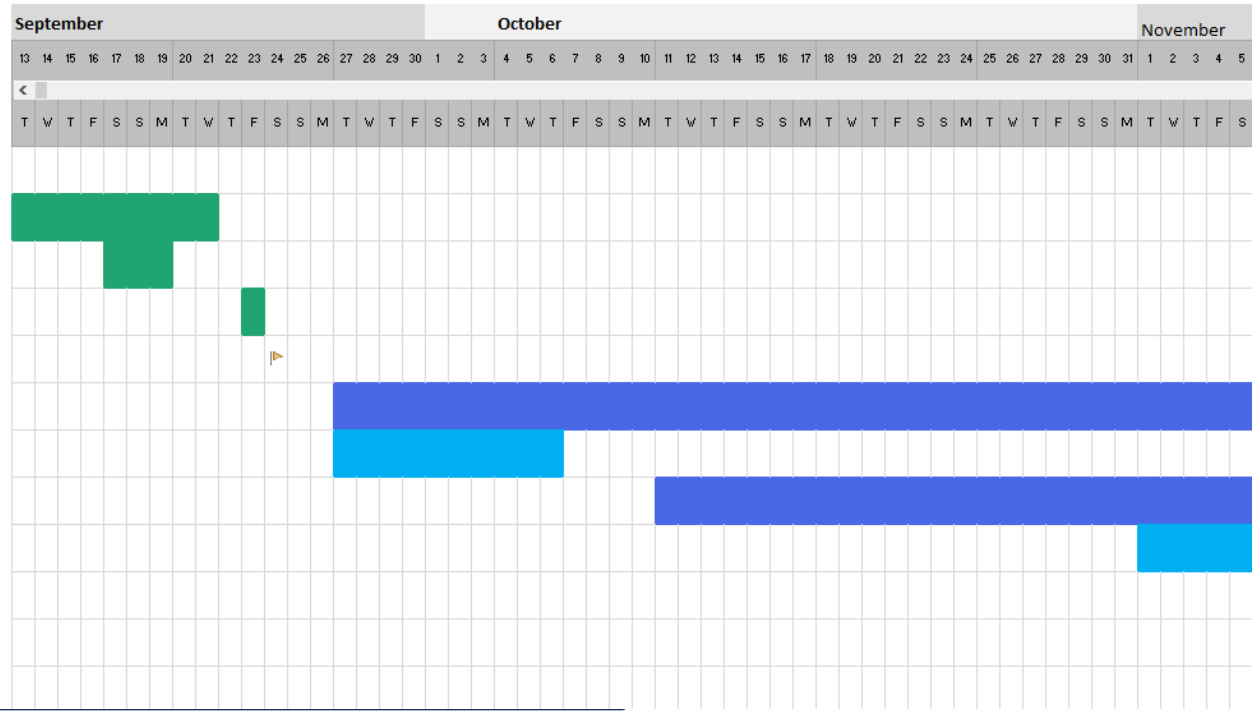
Project lead: Doug Leece

Project Start Date: 2022-09-12

Scrolling Increment: 1

Legend: On track | Low risk | Med risk | High risk | Unassigned

Milestone description	Category	Progress	Start	Days
TITLE 1				
Literature Review	On Track	0%	2022-09-12	10
Create test environment	On Track	5%	2022-09-17	3
Catalog datasets collected	On Track	5%	2022-09-23	1
Preliminary loading of public dataset	Milestone		2022-09-24	1
Start writing anonymizatoin programs	Med Risk	0%	2022-09-27	90
Create statistical validation scripts	Low Risk	0%	2022-09-27	10
Create privacy search criteria/validation	Med Risk	0%	2022-10-11	55
Start developing adversarial injection processes	Low Risk	0%	2022-11-01	10
Start developing analyst surveys	Low Risk	0%	2022-11-15	10
Finalize testing environment	Goal	0%	2022-11-16	5
Begin recruiting participants	Low Risk	0%	2022-12-01	20



- Duration 6 months
- Start dates module dependant
- Timelines reviewed / revised every 14 days

THANK YOU

DLEECE@FIRSTFIRETECH.CA

REFERENCES

- *BCS. (2022) Academic Accreditation guidelines. Swindon: BCS The Chartered Institute for IT*
- *Bresniker, K., Gavrilovska, A., Holt, J., Milojicic, D. & Tran, T. (2019) Grand Challenge: Applying Artificial Intelligence and Machine Learning to Cybersecurity. Computer 52(12): 45-52.*
- *Elastic. (2022) Elastic Common Schema: Normalizing Your Data with ECS. Available from: https://www.elastic.co/what-is/ecs [Accessed 19 August 2022].*
- *Essex (n.d.) MSc Cyber Security Project. Available from: https://online.essex.ac.uk/wp-content/uploads/One-page-module-guides/Computing/CYSPROJ.pdf [Accessed 19 August 2022].*
- *ICS2 (2021) The Cybersecurity Workforce Gap. Available from: https://www.isc2.org//-/media/ISC2/Research/2021/ISC2-Cybersecurity-Workforce-Study-2021.ashx [Accessed 18 July 2022].*

REFERENCES

- *LaFerrera, M. (2022) Introducing Synthetic Adversarial Log Objects (SALO). Available from: https://www.splunk.com/en_us/blog/security/introducing-synthetic-adversarial-log-objects-salo.html [Accessed 6 August 2022].*
- *Manokha, I. (2018) Surveillance: The DNA of Platform Capital-The Case of Cambridge Analytica Put into Perspective. Theory & Event. 21(4):891-913. Available from: https://muse.jhu.edu/article/707015 [Accessed 19 August 2022].*
- *Rashid, A., Chivers, H., Danezis, G., Lupu, E. & Martin, A. (2021) CyBOK: Cyber Security Body of Knowledge. 2nd ed. United Kingdom: The National Cyber Security Centre Available from: https://www.cybok.org/media/downloads/CyBOK_v1.1.0.pdf> [Accessed 19 August 2022].*
- *Sommer, R. & Paxson, V. (2010) 'Outside the Closed World: On Using Machine Learning For Network Intrusion Detection', 2010 IEEE Symposium on Security and Privacy. Berkeley, Oakland Ca, 16-19 May. New York: IEEE 305-316*

REFERENCES

- *Splunk. (2021) SALO Documentation - SALO v0.1.1 Documentation (Git Hub) Available from: https://splunk.github.io/salo/ [Accessed 6 August 2022].*
- *Trizna, D. (2020) Security Detections on Windows Events with Recurrent Neural Networks. Available from: https://ditrizna.medium.com/security-detections-on-windows-events-with-recurrent-neural-networks-346d0b2738fe [Accessed 6 August 2022].*
- *Yichiet, A., Khaw, Y.-M. J., Gan, M.-L. & Ponnusamy, V. (2022) A Semantic-Aware Log Generation Method for Network Activities. International Journal of Information Security. 21(3): 161-177. Available from: https://doi.org/10.1007/s10207-021-00547-6.*