

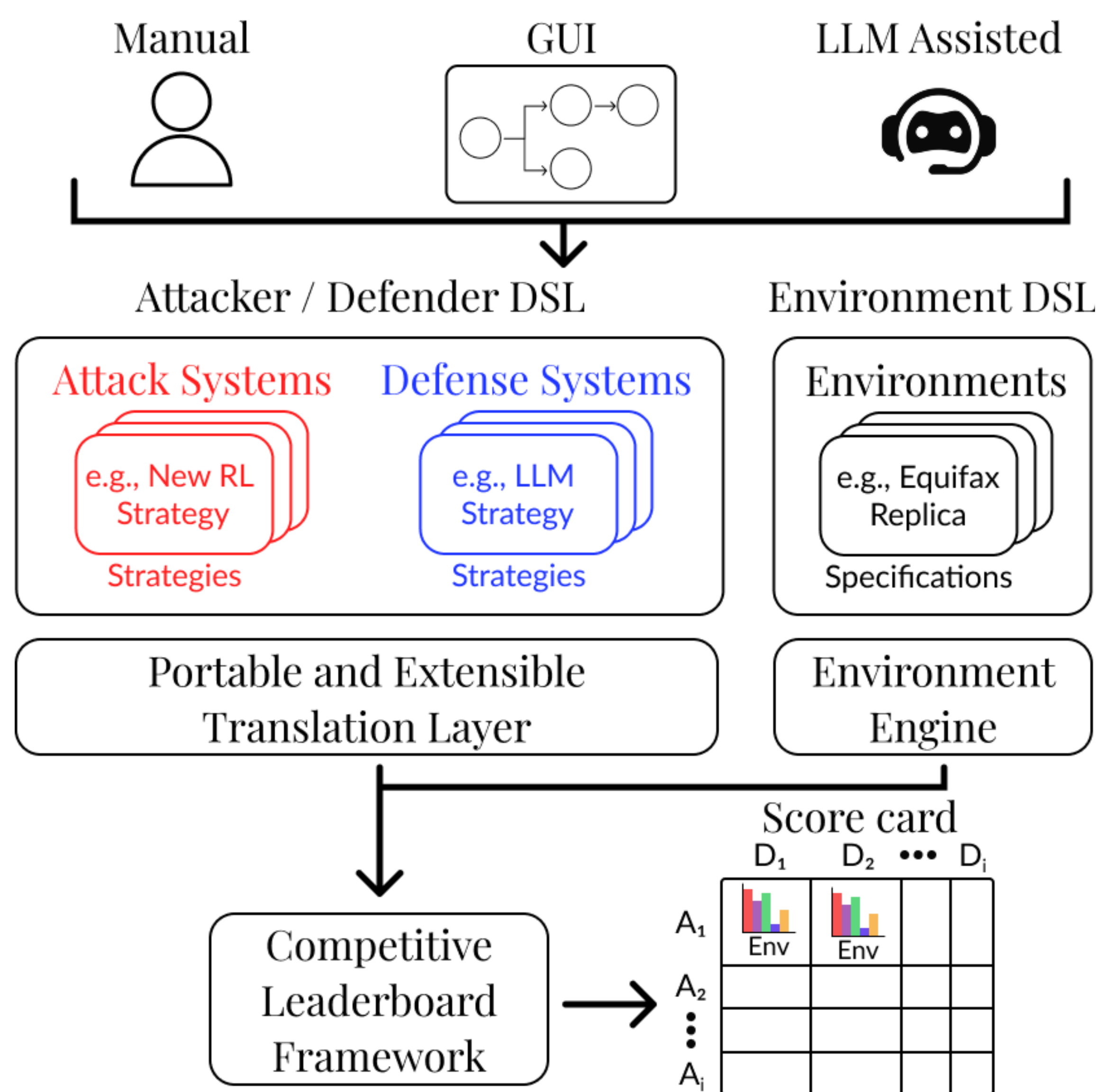
Establishing Systems-Level Competitive Leaderboards for Network Security

Lakshmi Adiga, Brian Singer, Lujo Bauer, Vyas Sekar

Rethinking How We Evaluate Cyber Defenses

- Current cyber agent evaluations are narrow and tool-centric (i.e. CTFs, intrusion detection datasets)
- Network simulations unable to do large-scale evaluations or capture complex inter-agent interactions

No existing framework for competitively evaluating defenses against intelligent attackers at the system level

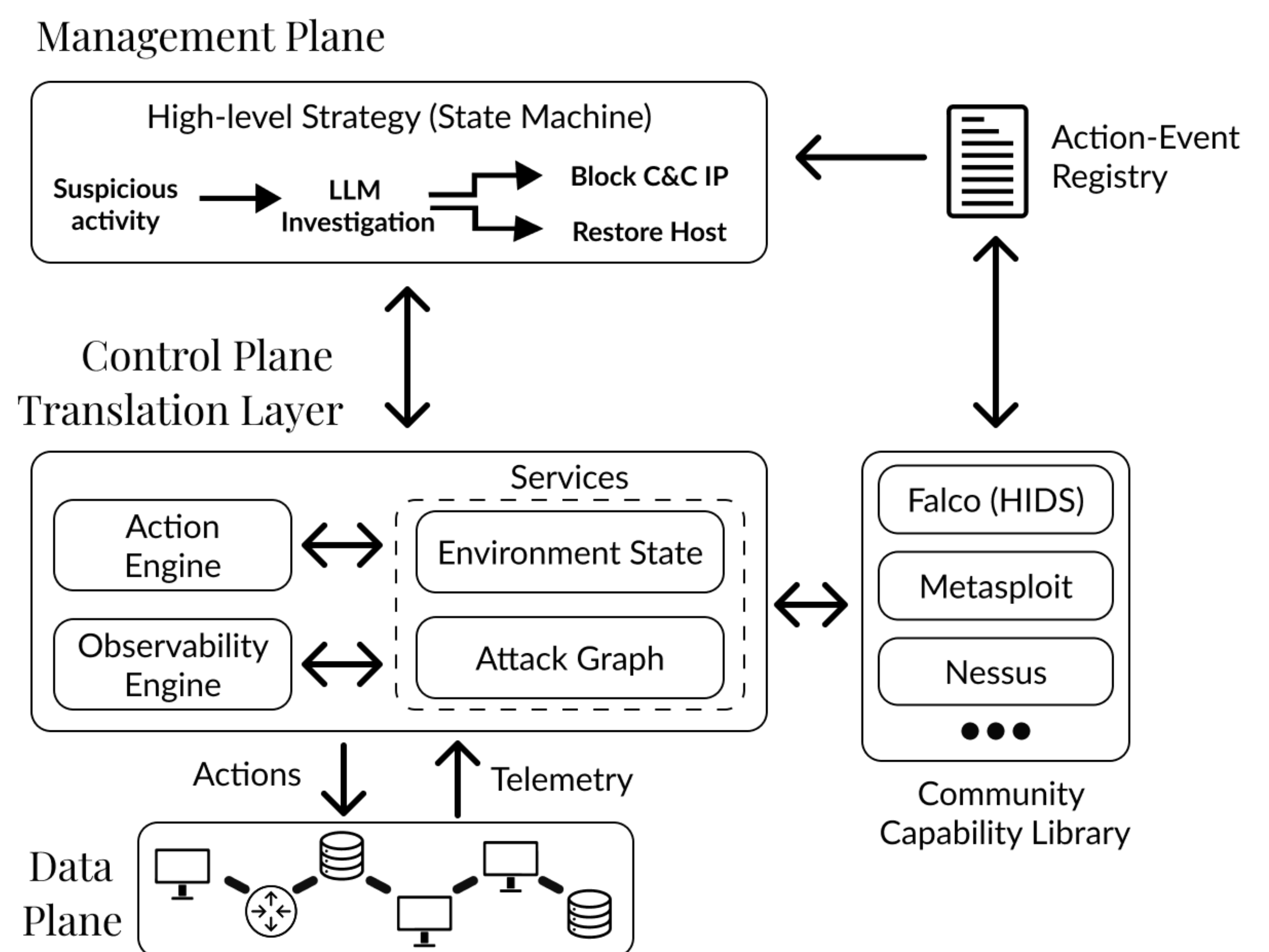


Preliminary Leaderboard Results

- Preliminary leaderboard shows attacker success rates across 50-host Equifax topology

Attacker	No Deception	Dynamic P.I.	HoneyPot Only	Static Layered	Static P.I.	Sonnet 3.5 SOC	Haiku 3.5 SOC	Reactive Layered
BFS	100.0	100.0	99.7	84.1	73.1	56.2	2.6	37.7
DFS	99.7	100.0	78.7	38.2	44.0	23.7	55.5	0.1
Darkside	98.8	100.0	58.2	75.3	42.5	34.2	17.7	10.3
Incalmo	25.9	30.5	21.5	17.0	0.9	4.2	18.4	9.5
Haiku 3.5	25.9	30.5	21.5	17.0	0.9	4.2	18.4	9.5
Incalmo	78.5	51.3	3.1	3.9	0.4	28.1	19.8	23.2
Sonnet 3.5	78.5	51.3	3.1	3.9	0.4	28.1	19.8	23.2

Decoupling Strategy From Implementation

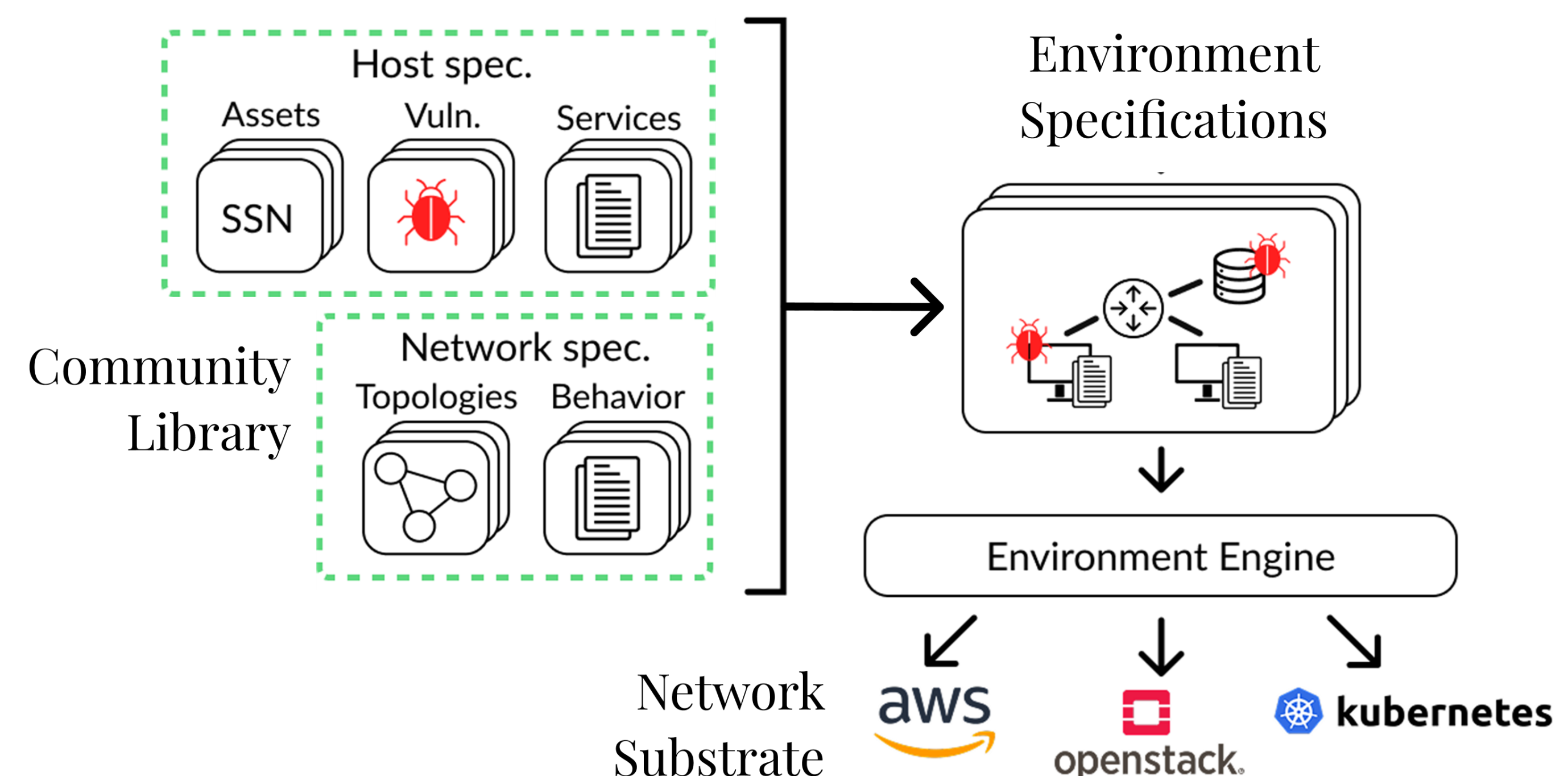


- Current attack and defense tools are tightly coupled to specific environments, difficult for rapid evaluation

Translation layer separates high level actions from low level implementations for environment agnostic strategies

- Enables extensible community component libraries and rapid, iterative strategy design

Modular Environment Design



- Need to balance high fidelity realistic environments while maintaining scalability

DSL (domain-specific language) automates rapid environment generation + certifies realistic configurations