DATA

Top Level Questions

How do we deal with the massive amounts of dynamically changing data that the cyberspace environment holds, including the identification of adversarial behavior?

Are the right data being collected to populate emerging models?

What tools and techniques from the VLDB/XLDB communities are relevant to this problem?

Anomaly Detection:

Comparison to Recent Behaviors

What are the key entities for which these aggregates and summaries need to be computed and maintained?

What aggregates and summaries need to be computed and maintained for anomaly detection to be effective?

Over what time period do these get maintained? (i.e., over what time periods are the phenomena roughly constant?)

How are they initialized?

Lessons from Other Fields

What techniques from other fields are used to detect harmful and/or adversarial behaviors in massive dynamic data sets? How are the key attributes of these fields and techniques similar or different from the attributes of cyber?

Models as basis for Comparison

How can we determine what is normal? How is normal different from benign? How can we detect abnormal behaviors? How does harmful behavior differ from abnormal? How to detect harmful behaviors that masquerade as normal?

How do we create, evaluate, and maintain effective models in a dynamic environment?

Maturity of Techniques

What is the maturity of each technique? Is it a:

State-of-the-practice in Cyber

State-of-the-practice in another field?

State-of-the-art in Cyber?

An identified open research question in Cyber?

An unaddressed/novel/too-hard research question?

What is the key barrier to use of this idea in Cyber?

Could the idea pay off in the short, medium, or long term?

Some Open Questions

How could we identify individuals (multiple-identities) in Cyberspace?
How could we identify coordinated actions by related individuals?
How can we separate hostile activities from benign activities by the same individuals?
How could we identify preparatory activities (for an attack) by individuals or groups, especially when they occur simultaneously with legitimate activities?

Evaluation

How do we know if a model is effective?
What criteria must a model satisfy to be useful? Credible?
What baselines are appropriate for evaluating a model?
How do we evaluate the value of a set of models? The incremental value of a new model? The cost of adapting a model, in terms of user understanding and acceptance?

Practicalities

What are the computational limits on real-time model evaluation?
How are models initialized and maintained?

How do we evaluate models that use attributes that aren't regularly computed?
What false-positive rates are acceptable?
How do we communicate the model structure to users so they can understand what the model is telling them?