**Questions on Webpage**

1. How do we deal with the massive amounts of dynamically-changing data that the cyberspace environment holds, including the identification of adversarial behavior?
2. Are the right data being collected to populate emerging models?
3. What tools and techniques from the VLDB/XLDB communities are relevant to this problem?

**Focus on cyber—data—models—novel/innovative/crazy ideas**

**Models vs building**

**Model vs use**

**Data=query--models**

1. Searching huge unorganized data sets is there any purpose on creating bigger haystacks. Techniques to get useful information. Practicality. How to extract useful data?
2. Ammonization how it would be useful, is that theoretically possible.
3. Connection problem. Large sets of data, 6th degree problem.

* Fundamental trade off. Ideas for trade off between decisions on streaming data on what is relevant and the challenges on what we don’t know is relevant based on context.
  + Assume models for things we know about that are relevant and what we don’t know are relevant yet.
  + Some data has more leverage than others.
  + **Outliers vs. inliers**- plots and distributions, but looking for bad guys in context we want the inliers. The first thing we may look at is an outlier; really want to catch those who do this on a regular basis. They have learned how to do it and are successful at doing it. This is based on the data we can collect though. Think about the things that people do not think about.
  + If we can identify the less voluminous data set that has higher signal to noise ration this will be valuable.
  + Assumption that needs to be validated
* More entities or more attributes
* We use data to help develop models and given the model detect instances.
* **Less data is good, more data is better**
  + From cyber we have uncertainty and no guaranteed connectivity. I have access, that is the only thing I have to worry about. More data is better, we trying to understand the behaviors. And that information comes in all the time and is inconsequential and we need to be able to analyze this data. Analyze data beyond the domain that we are looking at so we establish a context to ask new questions of data and data can give us questions we can ask.
* Models building vs. model use.
* What is the optimal data that the customer wants to answer, the questions that they want answered with high certainty. That may allow us to approximate the optimal data.
  + Wrong questions-- What are the categories of questions we expect to ask of the data? We need an adaptable system. The current system can only ask hard selectable questions
  + Certain model we are going to start with, we have things we can assume we want to ask, such as inliers. Augment the data that we collect, data about data, and how it can be used in the existing way that we use it today. End up wanting to open the aperture wider than we currently do.
  + What samples are we sampling today, and this might contain relative parameters. Our challenges require different data.
  + Data collection and analysis is to catch bad guys, **someone starts good they go bad**. Digital footprint, the analysis of current footprint is more important than a historical footprint. Which will build up in 2 or 3 months.
    - I modeled you in the past, see your money, and then see an anomaly in actions I know I need to watch you.
    - That assumes that you the bad guy is careless in changing actions
* In the end we should accumulate data and then get rid of the data not needed
* Part of the issue on how to go forward, we don’t get rid of asking specific questions but we add another axis of adding new questions
* Smaller scale, I keep everything, because you find something interesting and then you look back and see it. **Central accessing vs. distributed**?-Paradigm
* Computation to data not data to computation
* Massive sets of data—storing a lot of data centrally be able to look back historically to identify new patterns of data that can be applied to streaming data.
  + Sleeper actions added to it to make it malicious a lot can be done there with data capture you need historical data to see the instance of the sleeper.
* Mathematics compressive sampling- sample into that part of the parameter space, just to know about unknown
  + **Compression and encryption**
  + Speed accuracy trade off
* Take previous stuff, high level abstract model, but then we lose the detail. **How does something I have seen before become relevant based on what I just learned?**
* Assume anomaly detections need more data.
* What is the right amount of information that we need to do what we need to do?
  + In response to more data vs. less data argument
  + Need to characterize the data where it is locally relevant
  + Not all ink virus pushed out everywhere, because than can be solved everywhere. Look across domains so you can understand across entities. Can get down to causal events.
* **Connection between models and data, universal data applies to all models. Characterize data to partition data from global data and global data to local data and local models.** 
  + **Models are data and data are models**.
  + Models should be able to be decomposed
  + Calculus of models
  + How do we push context to edge because context is data too
* Malware strategy: do not push out malware everywhere. Artifacts may not appear in every domain
* Logs are most useful thing that it detects
  + For understanding on how that attacks happen and seeing if it happens in the future.
* What is the optimal data?
  + Offense and defense, pretty good at offense. Improve situational offense. The other side is trying to move to architecture in hardware and software. Layer analytics. The ability for analyst to see behavior and the model be able to find that optical device.
    - Need new analyst who is an expert in 5 fields, but does not scale to needs of the nation. Need to make process is the big research challenge.
* Proactive experiment to change context to look for consequences in the data stream. Usually gets constrained in policy.
  + **Sensor network**: In context of understanding what the sensors are, we interact with environment, we can sometimes perturb it and see what it does based on that. Make a linkage to get my next set of questions.
* When we pull data centralized. Only one person can see what I am searching for, but if I do it remotely everyone can see that I am searching. Confidentially on these sensors.
  + When no longer passive, then I have no reason to trust. Try to make it trustworthy, but if interactive, it is not trusted. How do you maintain tradecraft? How do you dynamically adjust and maintain tradecraft.
* Make bad guys do things more complicated lets us track it easier and has a better chance of them screwing it up.
* Data collection is passive or it is in trouble, but should consider it is something we can shape based on our actions. We do something to prompt a response.
  + **Active probe**
* What is the optimal data?
  + Immediacy of action on detection of cyber activity. Do not know the affect of your work. The action cycle in cyber is immediate. Actions have few degrees of freedom than the data.
  + What data do I need to set the actions.
* Actively probing attacker.
  + Probe attacker by actions made
* Probing and optimal data are related. Not stopping acquisitions of data at some planning table where someone else takes action
* Astronomers done—large data sets, different wavelengths, moving objects.
  + Distributed data, orphan basket
  + Locations of the data sets so they do not physically coexist
  + Moving object issue
  + Simulate images of sky, universe. **What is the role of simulations?**
    - in particle physics it is 100 simulations for one real thing
    - in astronomy it is 1 simulation for the universe
* Cyber issue—find the alien spaceship that if they discover Earth could come and harm us
* Create a team that inserts into data artifacts, real data from some set of feeds for academics to work on. They insert things to see if academics are finding.
* **Turn the victim into a honeypot**
* System that learns as it is attacked. **Automating responses**
* Crowd sourced red team—rewarded based on what they do, feed them ideas see what they adopt
* **We want to detect something that we have never detected, can we do this automated**.
* Most of our models are even driven. We should think in terms of classical physics. Somebody validate statistical mechanics of differential equations

**Suggestions for techniques and data sources. (New crazy idea)**

1. Data Source: Keystroke Dynamics
   1. Used to build models
   2. Used to characterize an individual
   3. Can possibly detect new behavior indicating when an individual goes bad
   4. Record use of networks as well
2. Data Source: All internet telemetry, DNS data traffic, whois, traffic, trace roots, zone filters, routing tables, tlds, control plane traffic.
   1. Bad guys will have to have preloaded their data centers with material they want.
   2. Domain registrations, behavioral aspects, dot registrations on a day. See indications of new threat
3. Data Source: Is there a correlation between threats faced by consumers vs. threats on the government?
   1. Is there any value in harvesting telemetry data from consumer elements to spot new adversaries?
4. Data Source: Data trails that analyst’s leave. Data that analysts generate while doing their jobs.
   1. Data will be shared in order to help other analysts do their job.
   2. Behavior analysis
   3. Would analysts be comfortable with this?
5. Data Access Model: Research to get access to data. Immediate model.
   1. Share model instead of data.
   2. I can’t give you my data, but you can give me your analytics to run on my data
6. Randomize response to attacks and study the mutations. Bacteria theory. Collect data on responses that work.
   1. Problems: feasibility. How can we automate this? How can we even build this?
   2. Malware itself has already developed randomized attacks
7. Data Source: 100$ brain wave reader. Ted Conference.
8. Do we really need modeling? Should we look directly at the data?
   1. Millions of people could have access to it, could we dissolve complex models into simple problems
   2. Complex, directly learned models from massive data that aren't necessarily understandable
9. Using visualization as a method of seeing things in the data. Dyslexics can be especially good at this.
10. Screening data - being able to assess temporal value of data, forensic when building models and when to throw away data. I want a better way of determining whether to throw away or retain data.
    1. continuous function for retaining data
    2. reimagining the data that we collect from traffic centers and networks
    3. redesign of IOS summary
11. Divide and recombine (comprehensive analysis which does not lose important information)
    1. not random samples
    2. collections of subsets based on value of conditioning variables
    3. recombine the results
12. Create a data set that can currently compute models
    1. compare models, find similar behaving models, and compare against own sensors
    2. use this to find anomalies
    3. data set of derivative models from network flows correlated with existing sensor feeds
       1. HMMs
13. Prevention, Pro-action, and Prediction
    1. Prevention - create a small configuration/checker that will go around to different machines to see if machine's configuration is compliant or not
    2. Pro-action - How do I establish a cryptographic analog for computer operation?
       1. Every individual has a lockbox with individual keys
    3. Prediction - Hierarchical spatio-temporal model applied to streaming for pattern discovery
       1. anticipate when to put on the brakes
       2. eyewitness problem - different takes on the same thing
       3. perhaps each take can agree on a core fundamental concept
14. Cyber MAD policy
    1. Could we create a MAD theory that applies to CyberSecurity?
15. "Crowd-computing" based data acquisition of user host data, internet scale, 1MM++
    1. like the DARPA red balloon challenge, but avoid gathering data from/about children
    2. download of a host sensor to gather data over a significant period of time
16. Longitudinal data of host/server "vulnerabilities" via a long term scan
    1. analyze what host/servers change over time