**Data Track Discussion—10:30 -11:00 Wednesday, August 18, 2010**

**Ted Senator Presentation**

* Data Issues
	+ What is optimal data?
		- “The unreasonable effectiveness of data”- article from Google people
			* Models don’t always help as much as data
				+ Study showed the amount of training data rather than sophistication of model let to greater performance
				+ Google like this because they are based on tons of data
			* More data over fancy model
		- Need both detail and broader context
			* Hierarchical data and hierarchical models
	+ Adaptive approach/moving target
		- Exploration Exploitations tradeoff
			* Learn models as the world is changing
			* ASAP learn when someone is doing something bad
		- Soon most data streaming only
		- System needs to learn and adapt
		- Continuous need for new data
	+ Detecting outliers and inliers
		- “Inliers”: need to change the representation
		- Do we want sensors that target something specifically or passively collect data
	+ Passive “sniffing” vs targeted sensor design
* Models and Data
	+ Take computation/analysis to the data
		- Much too much data, must take computation out to data
		- Decide at collection point/sensor what we need to throw out what to keep
	+ Data—models, two sides of the same
		- Models can be data and data can be models
	+ Massive sets of models are data
		- Use data to build/populate models
	+ Detect signature of good actor goes bad
		- Need historic trends as well
		- Longitudinal data
* Central vs Distributed Data
	+ Central: advantages in decisions making
	+ Distributed: tied to sensors/organizations
		- Data is heterogeneous
	+ Technical aspects, geo-location
		- Large data sets geo separated, sharing—in astronomical case
		- Astronomical-cyber is harder than astronomical because it fights back
	+ Can you do on demand sharing?
		- Data in different places that needs to be shared.
		- Many policy and technical issues with this
	+ How much analysis at the sensor?
		- Data reduction/latency
		- Only simple models at the sensor
		- Is the relevant propagated?
* Harvest Novel Data Sources
	+ Crowd source, games
	+ Internet telemetry
		- Trace routes, router tables
	+ Simulations
		- Red-team, blue team: insert synthetic events into real data stream and test recovery rate
	+ Correlations between data sources from multiple sensors/views of environment
* Book Citations
	+ The Shadow Factory
	+ The Watchers
* Questions
	+ Timing- What if we get the data too late to act on it?
		- Did you take into account the delays between when the data is taken and analyzed?
		- Sensor network to understand adversary-Intel strategy
		- Is there similarities between MS and the government?
			* Was the MS infrastructure compromised? Not yet?
		- Get one shot-make sure that it is not catastrophic
		- Counter terrorism-partial
	+ How much control does the enemy have over the data?
	+ If we have sensor network in place, how do we know robustness if it were attacked?
		- Could a sensor network detect if it were being attacked
		- Would it only take one attack but not the second?
	+ Data needs the models, and models needs the missions
		- Data feeds the models
			* Build, made, trained
			* Populate for detection
			* Explorations vs. exploitation
	+ How do you permit automated sharing of data?
		- Convince people to share the data?
		- How do we share horizontally?
		- We want to use data for model building requires different databases are infeasible to scale?
		- What needs to be shared based on if building the model or deploying the model?
		- Federated search vs. Federated Discovery
			* Search may be possible
			* Discovery is impossible; scale of search is too large
			* Astro is doing federated discovery
	+ Updating models?
	+ Improvement with human data added to large models
		- What is that key human data to be added to improve the model?
			* Couple of labels
			* Could be added to models, it’s hard to decide when I should go ask someone who has more access for more information
			* But can the analysts tell?
	+ How is labeling done today?
		- Don’t want to mark it, they just want to do their job, so you just take what they have
		- Different analysts do it different ways
	+ Simulations
		- Correlations help with the labels/ground truth
		- Know ground truth, will help to move the curve between explorations and exploitations.
	+ Crowd base competitions as a novel approach to use the internet—sending it home to gather data.
		- Google, MS have crowd sourcing available
		- Turn it in a game
		- Does not have to be outside internet, could just be the groups or smart computer scientists and engineers.
	+ Is there a business opportunity with large source network, signature distributions, talk to company to get additional data for fee, get into an already existing info structure?