Computing on Masked Big Data
Jeremy Kepner, Vijay Gadeppally, Pete Michaleas, Nabil Scheer, Mayank Varia
MIT Lincoln Laboratory

**Challenges**
- Users
- Architecture
- Cross VM Side
- Supply Chain
- Files
- OSINT

**Current Approaches**
- Operators
- Weather
- Enrichment
- Ingest & Enrichment

**Vision**
- Hypervisor
- Escalation
- Privilege

**Big Data Veracity**

**Performance Trade Space**

- Compute Overhead
- Operators (1000)
- Weather (100)
- Data Integrity (0.1)

**Step 1: Mask Data in Database**
- Plaintext input
- Stored masked data

**Step 2: Mask & Compute Analytic**
- Pedigreed correlation calculation in plain text
  - $A = \text{CatKeyM}(T,[\text{'src_ip|000.' : \text{src_ip|999.'})].')$
  - $T,[\text{'src_ip|000.' : \text{src_ip|999.'})].')$

- Pedigreed correlation calculation in masked text
  - $A = \text{CatKeyM}(T,[\text{'log_id|100 log_id|300'}].')$
  - $T,[\text{'log_id|100 log_id|300'}].')$

**Step 3: Unmask**
- Tallies row IDs between vertices

- Computing on masked data (CMD) raises the bar on data in the clear
- Uses lower over head approaches such as deterministic (DET) encryption and order preserving encryption (OPE)
- Associative array (D4M) algebra is defined over sets (not real numbers); allows linear algebra to work on DET or OPE data

---

This work is sponsored by the Assistant Secretary of Defense for Research and Engineering under Air Force Contract #FA8721-05-C-0002. Opinions, interpretations, recommendations and conclusions are those of the authors and are not necessarily endorsed by the United States Government.