The Applicability of HPC for Cyber Situational Awareness

Leslie C. Leonard, PhD

August 17, 2017

Outline

- **HPCMP Overview**
- **Cyber Situational Awareness (SA) Initiative**
- **Cyber SA Research Challenges**
- **Advanced Framework**
  - HACSAW
  - Data repository
- **Call for Proposals**
- **FY18 – Next Steps**
- **Questions**

HPCMP Ecosystem

Users

DOD HPC

A technology-led, innovation-focused program committed to extending HPC to address the DoD's most significant challenges

DoD Supercomputing Resource Centers (DSRCs)

- U.S. Air Force Research Laboratory DSRC
- U.S. Army Research Laboratory DSRC
- U.S. Army Engineer Research and Development Center DSRC
- Maui High Performance Computing Center DSRC
- US Navy DSRC

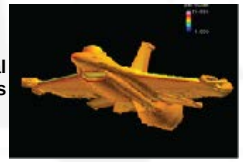
Networking and Security

Defense Research & Engineering Network (DREN)

Computer Network Defense, Security R&D, and Security Integration

Software Applications

Core Software



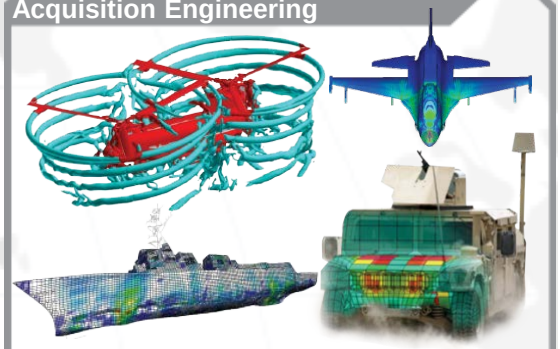
Computational Environments

Education and Training

HPC User Support

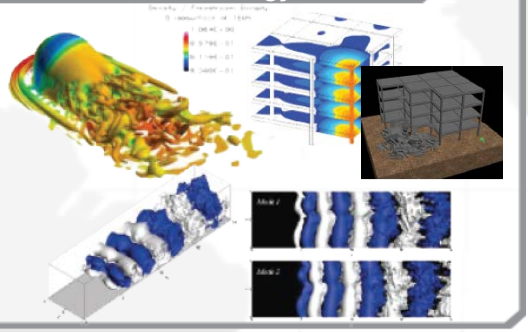
Results

Acquisition Engineering

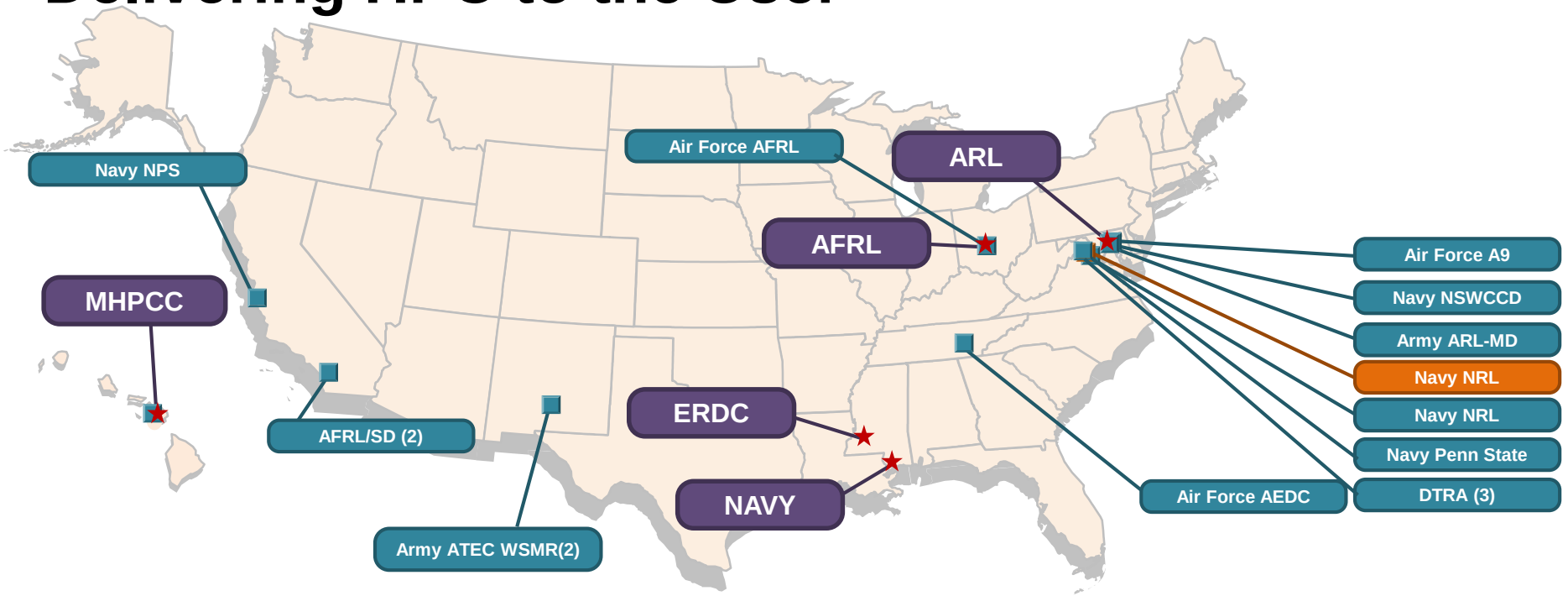


Test and Evaluation

Science and Technology



HPCMP Supercomputing Resources Delivering HPC to the User



DoD Supercomputing Resource Centers (DSRC)

- US Air Force Research Lab (AFRL)
- Maui HPC Center (MHPCC)
- US Army Engineer Research and Development Center (ERDC)
- US Army Research Lab (ARL)
- NAVY (Stennis Space Center)

Affiliated Resource Centers (ARC)

- US Air Force Research Lab, Information Directorate, AFRL-RI
- US Army, Space and Missile Defense Command SMDC
- Naval Research Lab (NRL)
- US Navy, SSC-SD

Dedicated HPC Project Investments (DHPI) FY13/14/15/16

- DTRA (2016, 2015, 2013)
- US Air Force, AEDC (2013)
- US Air Force, AFRL (2014)
- US Air Force, AFRL SD (2013, 2013)
- US Air Force, A9 (2016)
- US Army, ARL (2015)
- US ARMY, ATEC (2017, 2014, 2013)
- US Navy, NPS (2015)
- US Navy, NRL (2014)
- US Navy, NSWCCD (2013)
- US Navy, Penn Sate (2013, 2017)

11 of 17 DHPIs (FY13-17) operate at Above Secret

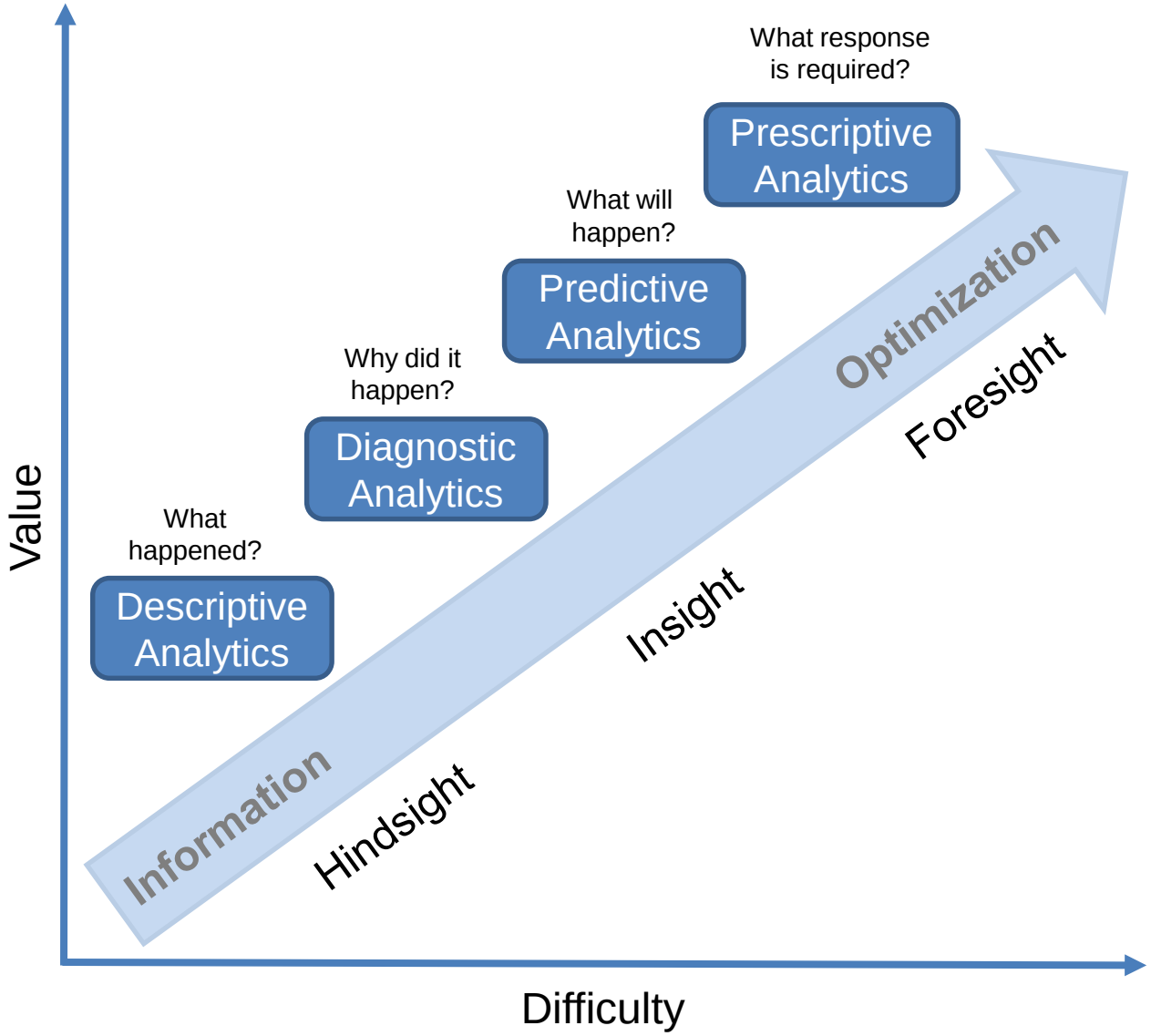
Cyber Situational Awareness Initiative

- **ESG-tasked initiative to, “examine the applicability of HPC to cyber situational awareness (SA)”**
- **HPCMP is well-positioned to leverage HPC systems to address complex cybersecurity problems**
  - World-class computational resources leveraged by the RDT&E community
  - Leading-edge software applications for computational analysis capabilities
  - National research and engineering network – DREN/SDREN
- **Multi-disciplinary, multi-year project leveraging expertise from HPCMP (e.g., Security, Networking, Centers, Software Applications) and external collaborators**

Research Challenges

- **Volume and complexity of cyber threats**
  - Require “real-time, automated” responses
  - Multi-dimensional threats
- **Needle in a haystack dilemma**
- **Hadoop and Spark vs. HPC**
- **Data**
  - Underutilized data sources
  - Data quality
- **Security analytics deployment**
  - System requires extensive configuration and/or tuning before it is usable

Stages of Analytics



Cyber SA Goals and Objectives

- **Explore current HPC and cyber SA intersections**
  - Understand how data analysis software performs on HPC
- **Conduct a formal analysis of cyber SA data sources**
  - Assemble raw cyber data streams and associated ontology
- **Perform an Open Source Software (OSS) compliance review**
  - Review all OSS components in order to validate compliance with requirements
- **Establish a data repository and HPC processing pipeline**
  - Ingest, parse, and store DREN data feeds
  - Establish & maintain a workflow model for HPC analytics
- **Support cyber SA and HPC analytic collaborators**
  - Collaborators perform data analytics studies against static datasets
  - Benchmarks are performed to compare HPC solutions with traditional non-HPC Solutions
  - Collaborators report and document findings

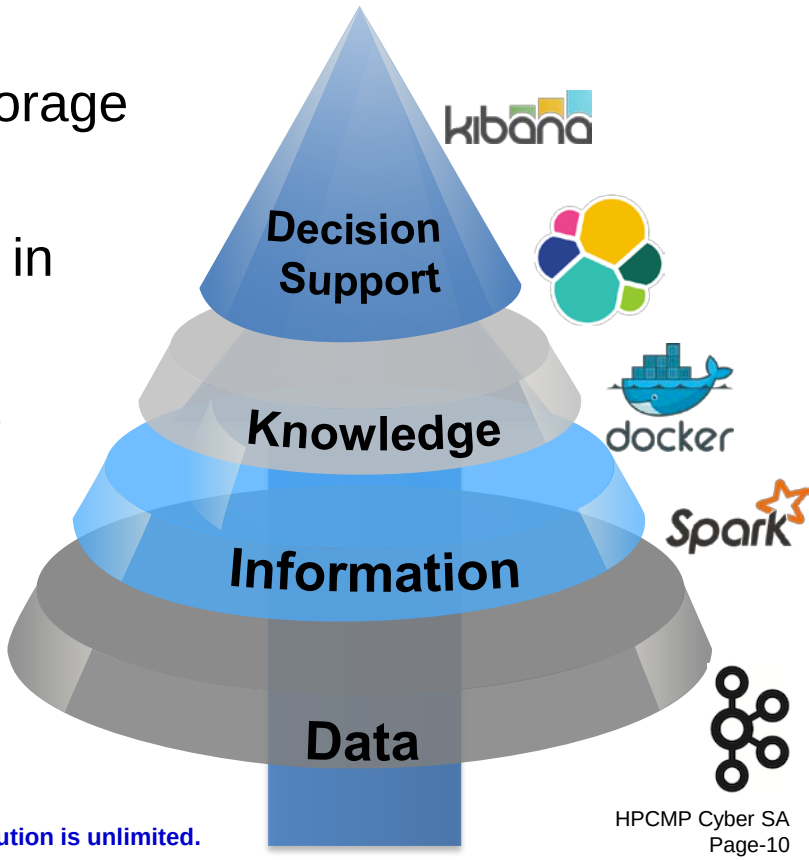
Advanced Framework for Cyber SA

- **Purpose:** To serve as the HPCMP cyber SA framework and facilitate the development and transition of data driven analytics
- **Project Name:** HPC Architecture for Cyber Situational Awareness (HACSAW)
- **Proving ground for novel ideas, algorithms, and approaches suitable for large scale execution in a dedicated HPC environment**
- **Reduce barriers to real world enterprise cyber data and computational resources**
  - ✓ HACSAW API v1.2.0 for quick data access
  - ✓ Container environment with common pre-installed and configured data science tools
  - ✓ HPC system dedicated to cyber SA data analytics
  - ✓ Documented workflow environment for collaborators

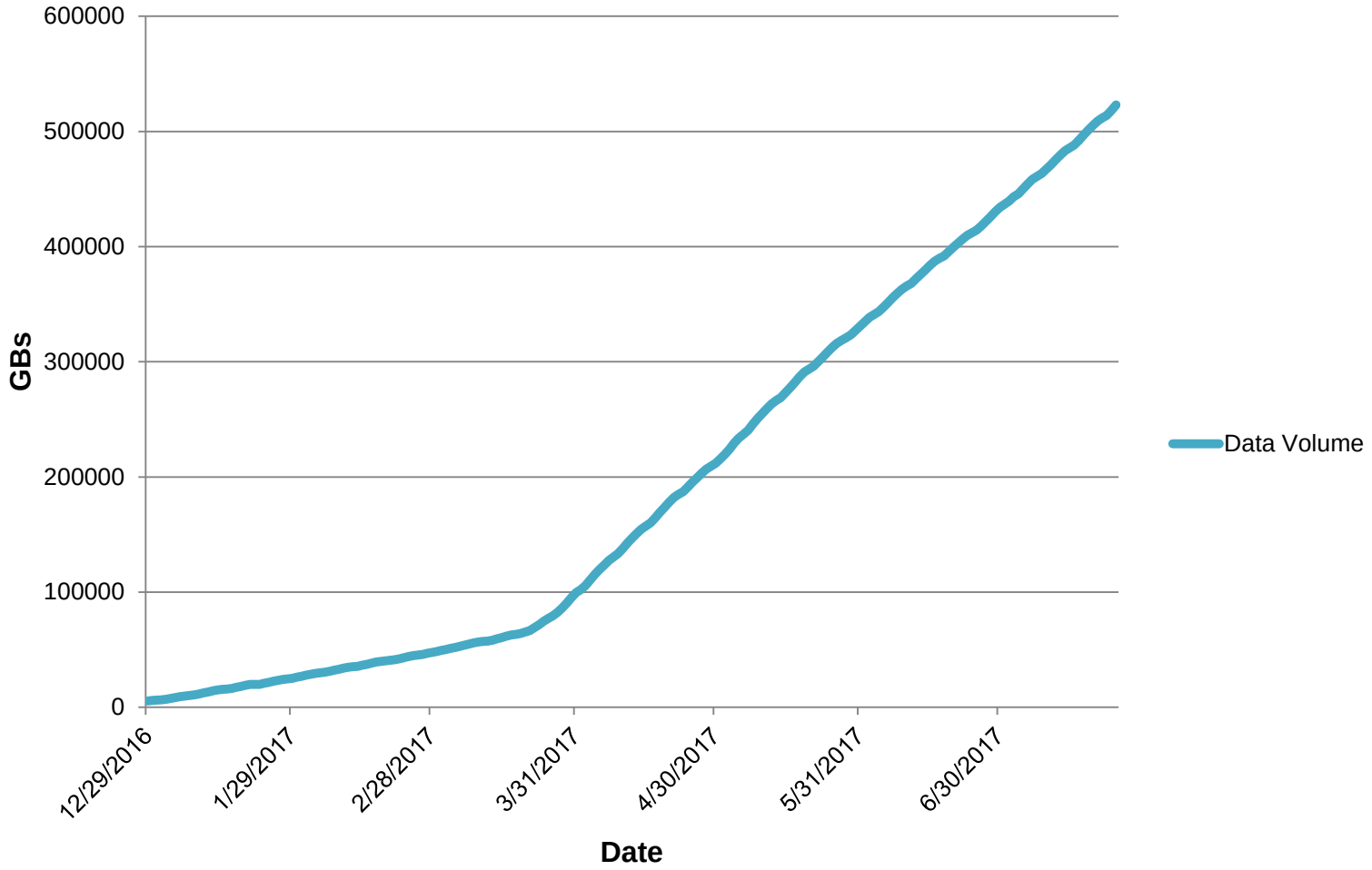
Data Repository

Most comprehensive cybersecurity data set available to DoD R&D community

- ✓ Collection of data sources from Internet Access Points (IAPs) to regional Service Delivery Points (SDPs) to the host-level
- ✓ Non-anonymized data
- ✓ Processing pipeline with redundant data storage and controlled access
- ✓ Rapid acceleration and exponential growth in size and complexity
- ✓ Proven open source, big data technologies for data enrichment throughout processing pipeline

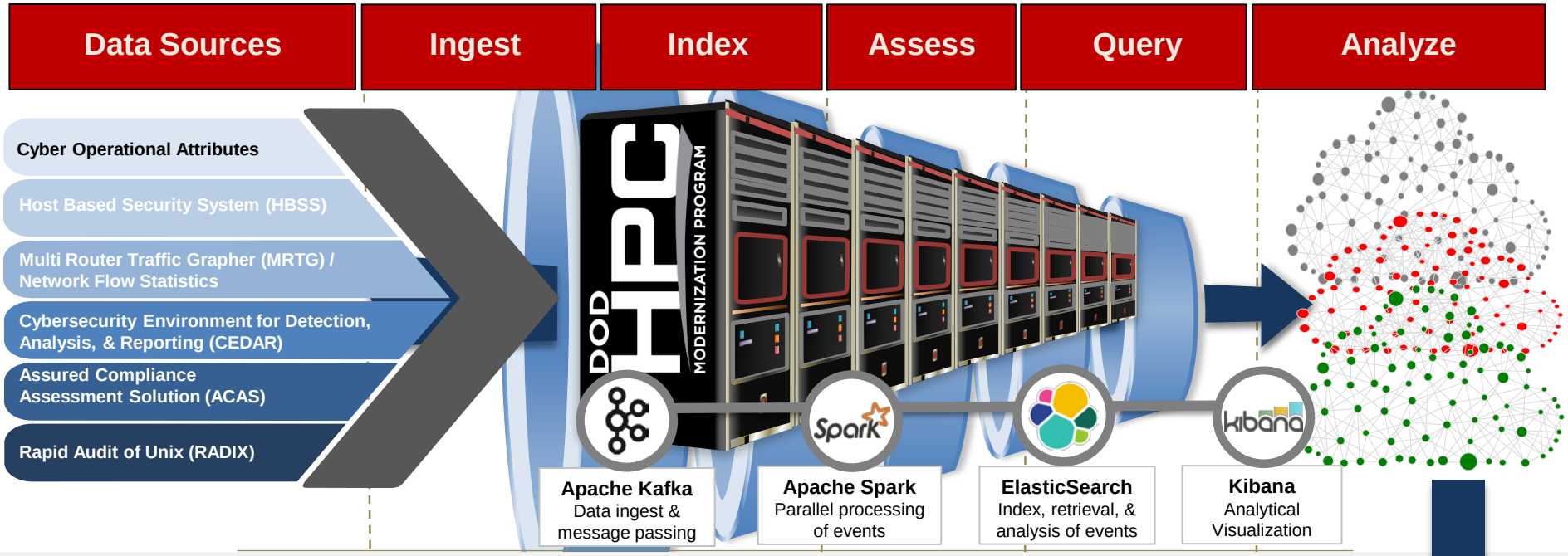


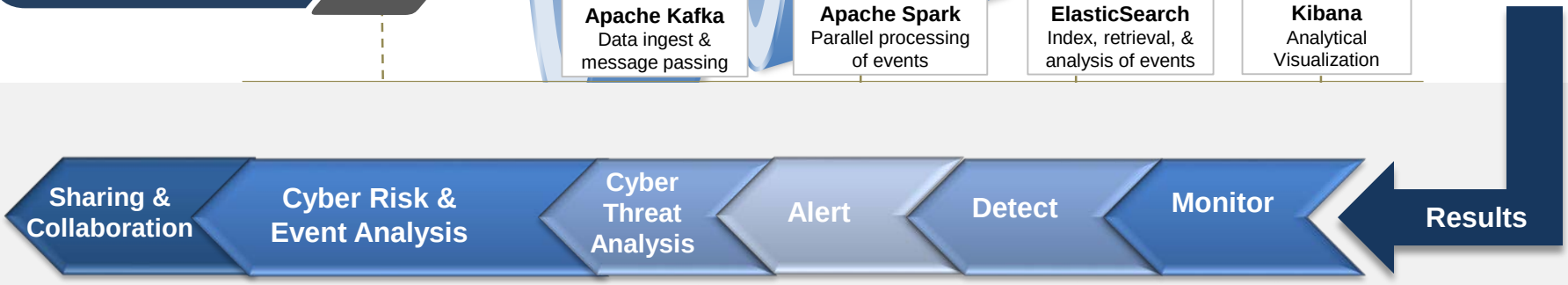
Total Data Feed Collection ~520 TBs



*Processing ~8 billion events/day, ~92k events/second, ~3.5TBs/day
as of 25 July 2017

HACSAW – The Big Picture





Cyber SA Mission Essential Tasks (MET)

HPC Development Environment

- **Hardware (virtualized)**
  - 36 cores 3.2 GHz
  - 117 GB user accessible memory
  - TB's of user accessible storage
  - Computational equivalent of one Topaz node
- **Software**
  - Use a standard analytical environment by using a Docker container
  - Supports repeatable experimentation
  - Includes common software, Spark, Dask, Pandas, Jupyter
  - APIs to access data via Python module
  - All software in container is compatible with the HPC environment
  - Additional software can be added if it meets requirements

HPC Development Workflow

1. DATA EXPLORATION

Identify relevant data sources and its underlying structure, purpose, and usefulness. In this stage, collaborators will exercise APIs and ontology to be used in the initial analytic development.

2. INITIAL DEVELOPMENT

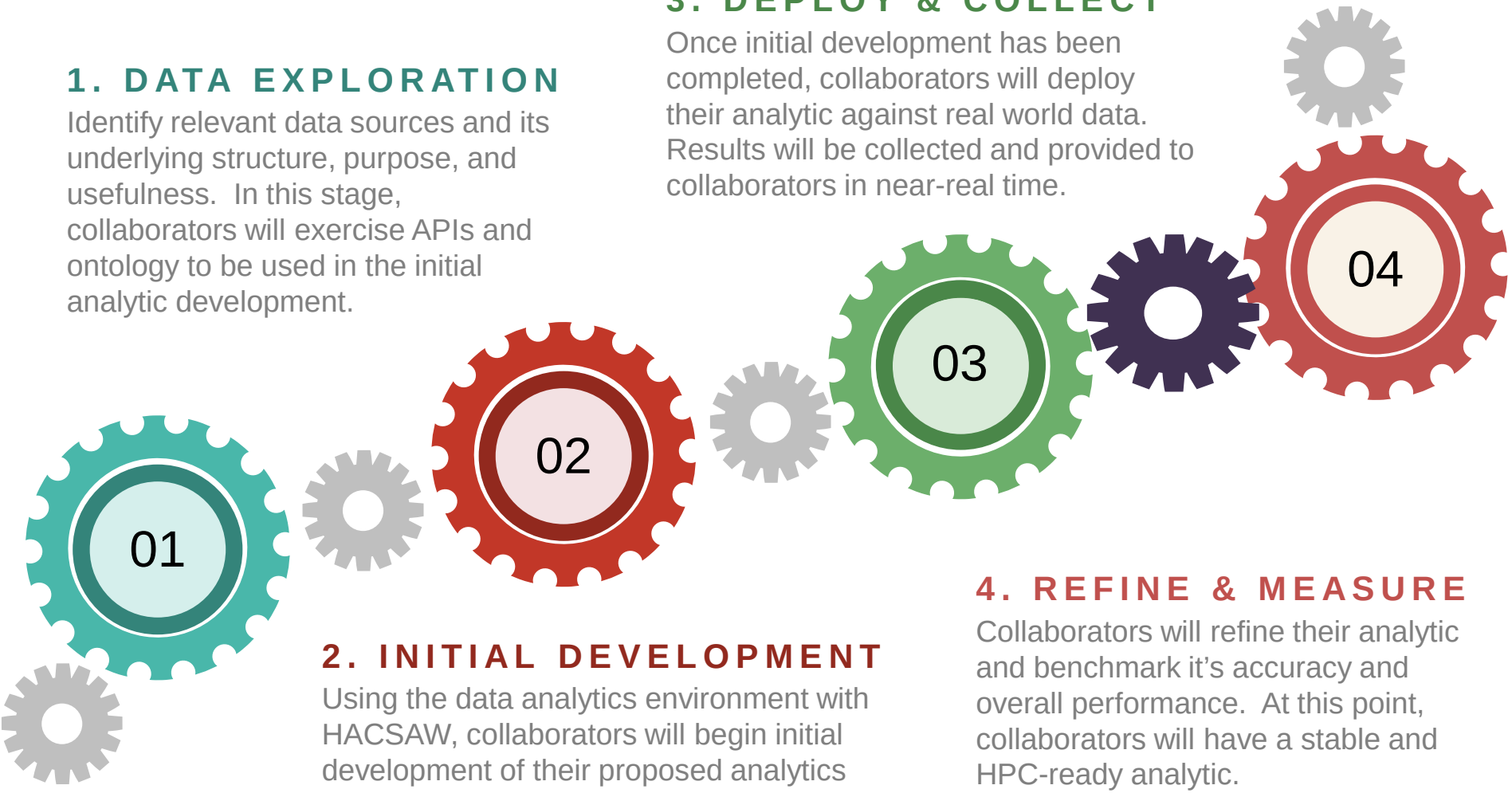
Using the data analytics environment with HACSAW, collaborators will begin initial development of their proposed analytics by working with real HPC data.

3. DEPLOY & COLLECT

Once initial development has been completed, collaborators will deploy their analytic against real world data. Results will be collected and provided to collaborators in near-real time.

4. REFINE & MEASURE

Collaborators will refine their analytic and benchmark it's accuracy and overall performance. At this point, collaborators will have a stable and HPC-ready analytic.



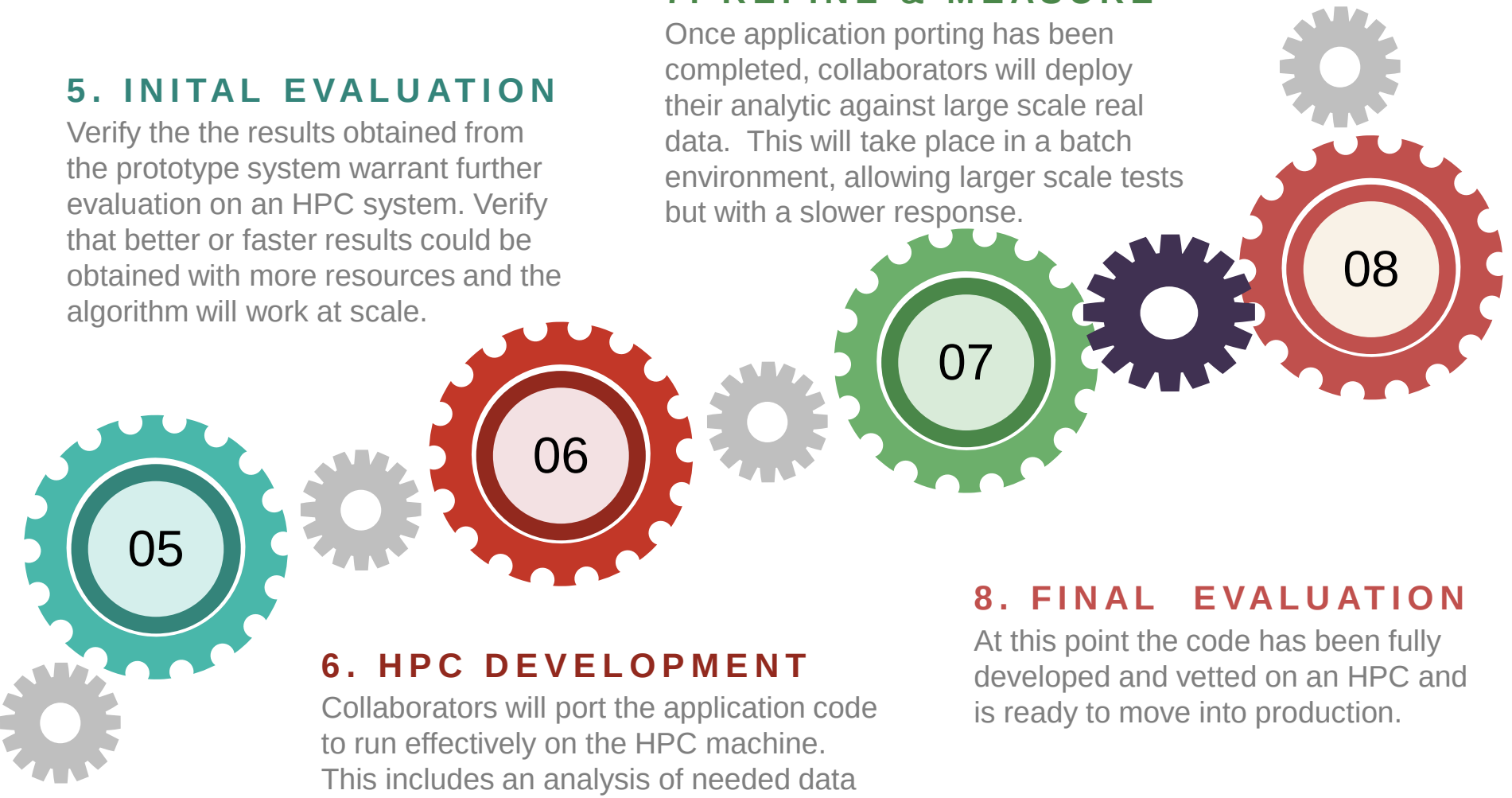
HPC Workflow

5. INITAL EVALUATION

Verify the the results obtained from the prototype system warrant further evaluation on an HPC system. Verify that better or faster results could be obtained with more resources and the algorithm will work at scale.

7. REFINE & MEASURE

Once application porting has been completed, collaborators will deploy their analytic against large scale real data. This will take place in a batch environment, allowing larger scale tests but with a slower response.



6. HPC DEVELOPMENT

Collaborators will port the application code to run effectively on the HPC machine. This includes an analysis of needed data and working with data owners to ensure data availability on the HPC.

8. FINAL EVALUATION

At this point the code has been fully developed and vetted on an HPC and is ready to move into production.

Workflow Walk-through

Example: Develop machine learning classifier (random decision forest) to detect anomalous browser User-Agents

- Data: HTTP logs, logs from existing pattern alerters
- Methodology: Create training sets based on existing alerter rules over various periods of time, and with varying training methods
- Expected result: Classifier will perform as well as existing anomalous user-agent detection alerters (95%>)

Workflow Walk-through

1. **Initial Development**
   - Docker container with provided software
   - Development on interactive system with real data
   - Data API accessed via Python module
2. **Deploy & Collect**
   - Train random forest using largest possible time periods on development system
   - Compare results with existing pattern alerters
3. **Refine Algorithm**
   - Evaluate methods and training data size
4. **Initial HPC Evaluation**
   - Ensure the developed application will transfer to the HPC
5. **HPC Development**
   - Improve algorithm scalability
   - Define data movement to HPC
6. **Refine Algorithm**
   - Evaluate methods and training data size
7. **Final Evaluation**
   - Perform training on HPC with large data sets
   - Deploy these classifiers on real time and historical data
   - Compare to existing pattern alerters
8. **Production**
   - Collaborate with HACSAW team to deploy classifier against real-time operational data
   - Create policies for periodically retraining data

Call for Proposals

- **Purpose:** Solicit proposal and work effort that yields results during a one-year effort that demonstrates potential for integration into the HPC Cyber SA operational environment and aligns with Mission Essential Tasks (METs)
- **Strongly encourage teams that span organizations**
  - Teams must include data science, cybersecurity and HPC expertise
- **Secret Security Clearance required for access**
  - No file transfers
  - Must sign document acknowledging data access restrictions
- **Important dates**
  - ✓ March 24: Call for Proposals
  - ✓ May 22: Proposals Due
  - ✓ July 25: Final Evaluation
  - August 25: Anticipated Award Announcement

FY18 – Next Steps…

- **Establish Board of Directors**
  - Provide oversight for HACSAW activities
  - Develop strategic direction and provide guidance
- **Refine HPC Cyber SA architecture**
  - Update HACSAW API, continuous HPC processing pipeline improvement, expand/refine data repository and software stack
- **Discovery, exploration and enlightenment**
  - Collaborators perform data analytics studies against static datasets that align with cyber SA METs
  - Benchmarks are performed comparing HPC solutions with traditional non-HPC solutions
- **Determine the future of the project**
  - Collaborators document and report findings
  - Major decision point regarding continuation or modification of project objectives

Conclusion

- **Multi-tiered approach** to ensuring a productive, secure, and trustworthy environment for the RDT&E and acquisition engineering communities
- **Cybersecurity investment** to increase the HPCMP's current and predictive understanding on the DREN
- **Continued engagement and collaboration** with Services and Agencies to leverage HPC, cyber and data science expertise for shared situational awareness

Questions?

Leslie C. Leonard, PhD
Cybersecurity Research Lead
Leslie.Leonard@hpc.mil

Abbreviations and Acronyms

TERM	DEFINITION
API	Application Program Interface
DoD	Department of Defense
DREN	Defense Research and Engineering Network
ERDC	Engineer Research and Development Center
ESG	Executive Steering Group
HACSAW	HPC Architecture for Cyber Situational Awareness
HPC	High Performance Computing
HPCMP	High Performance Computing Modernization Program
HTTP	Hyper Text Transfer Protocol
IAP	Internet Access Point
MET	Mission Essential Task
SA	Situational Awareness
SDP	Service Delivery Point