An Enterprise Approach to Evaluating Complex Systems Using Big Data Analytics

Ryan Norman
TRMC Initiative Lead for Big Data & Knowledge Management
PM, TENA SDA
ryan.t.norman.civ@mail.mil
So far, we have identified seven fundamental questions relevant to our domain that can be addressed by Big Data Analytic Techniques:

- Anomaly Detection – *Did something go wrong?*
- Causality Detection – *What contributed to it?*
- Trend Analysis – *What’s happening over time?*
- Predicting Equipment Function and Failure – *When will something go wrong?*
- Regression Analysis – *How is today’s data different than the past?*
- Data Set Comparison – *Are these two large data sets equivalent?*
- Pattern Recognition – *Are there any recognizable patterns in the data?*

These are not new or unique to T&E. So what’s the problem?
Worldwide Exponential Growth of Data

More information over the next two years than in the entire history of mankind!

<table>
<thead>
<tr>
<th>Name (Symbol)</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>kilobyte (kB)</td>
<td>$10^3$</td>
</tr>
<tr>
<td>megabyte (MB)</td>
<td>$10^6$</td>
</tr>
<tr>
<td>gigabyte (GB)</td>
<td>$10^9$</td>
</tr>
<tr>
<td>terabyte (TB)</td>
<td>$10^{12}$</td>
</tr>
<tr>
<td>petabyte (PB)</td>
<td>$10^{15}$</td>
</tr>
<tr>
<td>exabyte (EB)</td>
<td>$10^{18}$</td>
</tr>
<tr>
<td>zettabyte (ZB)</td>
<td>$10^{21}$</td>
</tr>
<tr>
<td>yottabyte (YB)</td>
<td>$10^{24}$</td>
</tr>
</tbody>
</table>

40,000 BCE cave paintings
Bone tools

3500 writing C.E.
Paper

1450 printing

1870 electricity, telephone

1947 transistor

1950 computing

Late 1960s Internet (DARPA)

1993 The web

2005 Mobile Internet (Smartphone)

2007 iPhone

2008 Social Web

Test & Evaluation Growth of Data

T&E Mission: Acquire data and discern into knowledge

Increased System Complexity

- Larger Test Footprints
  - 4-on-4 test flights (more systems per test)
  - Much faster weapons systems
  - Geographic separation not as effective as it used to be

- Demand for Shorter Acquisition Cycles
  - More concurrent testing
  - More real-time analysis

- Increased System-of-Systems Test Complexity
  - “Five Futures” (EW, UAV, NCO/W, DE, Hypersonics)
  - Integrated fleet (F-18E/F, E-18G, F-35, SM VI, UAV)
  - “Swarming” UAVs

Total Throughput: 7.5Mbps – 70Mbps+
Big Data / Knowledge Management (KM) Challenges & Needs

T&E Infrastructure Challenges:

- How do we conduct T&E of increasingly complex, data-driven systems?
- How do we enable more efficient & continuous system evaluation?

Need: A DoD-wide KM capability for T&E to help achieve better acquisition outcomes and reduce costs

- **Trusted processes** across government and industry that identify problems sooner rather than later
- **Accessibility** of knowledge & data to legitimate users
- **Discoverability** of knowledge & data obtained over time
- **Availability** of knowledge through common tools & technologies – including DoD T&E cloud solutions
- **Leverages** proven Industry techniques / practices

Big Data Analytics depends on effective Knowledge Management
The TRMC “Blueprint”: Putting Test Capabilities on the DoD Map

Defense Strategic Guidance

Acquisition Process

Service T&E Needs and Solutions Process

Annual T&E Budget Certification

Strategic Plan for DoD T&E Resources

DT&E / TRMC Annual Report

Risk mitigation needs
Technology shortfalls

Risk mitigation solutions
Advanced development

Requirements
Capabilities

TRMC Joint Investment Programs

(6.3 Funding)

Transition

Service Modernization and Improvement Programs

Acquisition Programs and Advanced Concept Technology Demonstrations

T&E Multi-Service/Agency Capabilities

DoD Corporate Distributed Test Capability

DRAFT UNCLASSIFIED – DISTRIBUTION STATEMENT A
Reference Number 15-S-2335; 08-14-2015
Realizing Improved DoD T&E Knowledge Management

1. Understand and Document T&E challenges & needs
   - *(FY12)* Completed Data Management for Distributed Testing (DM-DT) Study
     - Result: Developed functional requirements for T&E enterprise distributed Data Management
   - *(FY13)* Comprehensive Review of T&E Infrastructure report published
     - **Key Recommendation**: Use DoD cloud solution for T&E data
     - **Key Recommendation**: USD(AT&L) establish a DoD-wide KM capability for T&E to help achieve better acquisition outcomes and reduce costs

2. Execute proofs of concept that inform an enterprise approach to T&E Knowledge Management
   - *(FY14-15)* Joint Strike Fighter Knowledge Management (JSF-KM) project
     - **Goal**: Assess KM technologies and methodologies in support of an existing acquisition program
   - *(FY15-16)* Collected Operational Data Analytics for Continuous Test & Evaluation (CODAC-TE) project
     - **Goal**: Apply KM technologies and methodologies across the lifecycle

3. Develop investment plan that achieves strategic objectives:
   - Integrate T&E infrastructure into cohesive Knowledge Management enterprise
   - Modernize T&E practices & processes to leverage Big Data analytics techniques
   - Apply Big Data analytics tools & techniques to the T&E mission space
Example Framework View: Big Data Software Architecture

### Analytic Services
- **Quick-Look**
- **Real-Time**
- **Continuous**
- **Alerting**
- **Scheduling/Automation**
- **Analysis Tools**
- **Legacy Tools**
- **Generate Reports**
- **Audio/Video Analysis**
- **Build Queries**

### Data Analysis Packages
- **Scripting**
- **Streaming**
- **Graph-Based**
- **Data Mining**
- **Statistics**
- **Machine Learning**

### Data Services
- **Data Lifecycle**
- **Enforce Policies**
- **COO/DR**
- **Workflow**
- **Pipeline**
- **Ontologies**
- **Crawl/Index**
- **Tagging**
- **Versioning**

### Security
- **Authenticating**
- **Authorizing**
- **Access Control**
- **Enforcing Policies**
- **Enforcing Workflow**
- **Threat Detection**
- **Intrusion Detection**
- **Active Defenses**

### Query Engine
- **Structured Data Engine**
  - SQL Services
  - Working Sets
  - Tables
- **Unstructured Data Engine**
  - Key-Value Store
  - Filter
  - Sort
  - Summarize
  - Parallellize
  - Optimize
- **Structured Database**
  - Warehouse
  - Metadata
  - Replication
  - Schema
  - Resource Mgmt
- **Unstructured/Semi-Structured Database (Hadoop)**
  - Virtualized New Tools
  - Virtualized Legacy Tools
- **Abstraction Layer (Virtualization)**
  - Virtualized New Tools
  - Virtualized Legacy Tools

### Massively Parallel Tiered Computing, Storage, and Network Infrastructure
- **Existing Range Computing and Storage**
  - Existing Computers
  - Raw Files
  - Existing Range Databases
  - New Databases
  - Distributed File System
  - Remote Data Replication
  - Computing Resources

### Customization
- **User-Defined Analytic Plugins**
- **Customized Displays**
- **Customized UIs**
JSF T&E Infrastructure Needs Addressed by JSF-KM Project

1. **Data Warehousing**: Flight test data should be stored in a government facility to expedite data access & discovery

2. **Data Ingest**: Current DART Pod test data ingest is too slow to meet multi-ship quick-look and quick-turn requirements – examples: 2 on 2; 4 turn 2; 4 on 4 turn 4 on 4

3. **Data Access**: Test data should be available for quick-look analysis during mission debrief to inform decision making

4. **Video**: DART Pod video should be available for quick-look analysis during mission debrief to inform decision making

5. **Big Data Analytics**: Analysis capabilities need to proactively identify “unknown unknowns” and other anomalies impossible for a human to discern
Joint Strike Fighter Knowledge Management (JSF-KM) Test Concept

• DT & OT data storage in government facilities
• Collect / Store more precise data during OT
• Search / Analyze Edwards & Nellis data from any secure location
• Bring enhanced JMETC infrastructure to JSF T&E
• Apply commercial Big Data and Knowledge Management tools to DoD requirements
• Knowledge shared across JSF DT / OT T&E locations
• Scalable to other JSF T&E locations
JSF-KM Improvements to Existing T&E Capabilities

**DT Today**

- Data Ingest
- Raw Data Available
- Data Ready for Use @ LM
- Govt. Analyst Data Request
- Analysis

2 hours (per aircraft)

1 day

1 week

**OT Today**

- Data Ingest
- Raw Data Available
- Data Ready for Use @ (Govt)
- Govt. Analyst Data Request
- Analysis

1-2 hours (per aircraft)

10 minutes

4-5 hours

**With JSF-KM**

- Parallel Data Ingest
- Raw Data Available
- Data Ready for Use @ (Govt)
- Govt. Analyst Data Request
- Analysis

30 minutes (multiple aircraft)

30 seconds

90 minutes

Note: Numbers reflect single 2 hour flight mission
JSF-KM FY15 Success Stories

• Resolved test data time correlation issues
  – Time stamps in data files found to be corrupted post-mission
  – JSF-KM analysis tools were able to correct the time correlation issue
  – Without JSF-KM, at least five missions would have been re-flown

• Video available during post-mission debrief due to JSF-KM data ingest improvements from DART Pod
  – Existing tools could not process video in time to support post-mission de-brief
  – Without JSF-KM, there would be no flight video during post-mission debrief

• Discovery of avionics box issue during first night mission
  – Pilot and Analyst discovered problem from video data available 30 minutes after landing
  – Avionics Box was replaced before another mission was flown
  – Without JSF-KM, problem would not have been discovered for several days

Return on Investment has been realized before deploying any Big Data analytics capabilities
JSF-KM FY16 Success Stories

- **Reduced data profile time from 5+ hours to 47 seconds per Query**
  - Big Data tool enabled massive improvement to data profile generation
  - Without JSF-KM it would still take 5+ hours to perform data profile data runs

- **Identified 2 engines that consistently performed differently than the other 9 engines within the 92 data sets analyzed**
  - Discovered a faulty/noisy ground sensor and an unknown pattern within a known sampling rate abnormality
  - Without JSF-KM these anomalies may have never been discovered

- **Identified issue with ground sensor**
  - Found anomalous points and pattern within inconsistent sensor data sampling rates
  - Without JSF-KM these anomalies may have never been discovered
Collected Operational Data Analytics for Continuous Test & Evaluation (CODAC-TE) Proof of Concept

- **Background**: US Army Aberdeen Test Center (ATC) has ~30TB of underutilized data – including 20TB of in-theater operational data.
- **Goal**: Utilize Big Data analytics across multi-commodity DT, OT, and in-theater system data to discover “unknown unknowns” for current and future Army systems.
- **Use Cases**: Mine-Resistant Ambush Protected (MRAP) Theater Data; Camouflage Effectiveness.
- **Leverages**: High-Performance Computing Major Shared Resource Center and ATC expertise.

**Challenges being addressed:**
- Insufficient data science expertise.
- Current analytical systems inadequate for today’s data volume and velocity.
- Lacking tools and techniques for discovering unknown unknowns and conducting complex trends analyses.
CODAC-TE Success Stories
Phase 1

• **In-theater MRAP data collected, aggregated, and analyzed as resolution, complexity, and processing requirements increased**
  – Existing data analysis platforms and tools cannot process large enough datasets in a reasonable time or at all
  – Without CODAC-TE, ATC would not be able to scale to meet growing data sizes and processing requirements

• **Incorporated industry best practices for advanced data tools and analysis**
  – Approach allowed advanced data analysis tools to be available under rapid development schedule
  – Without CODAC-TE, using big data analytics would be cost & schedule prohibitive

• **Increased discoverability of multi-commodity life-cycle data and knowledge**
  – Phase 1 created an analysis platform capable of scaling to support breadth of data over time
  – Without CODAC-TE, available analysis tools could not load and process data across commodity areas or across all events in a system's life-cycle

Phase 1 completed on schedule and under budget
CODAC-TE Technical Accomplishments

Developed a platform to store and query data at large scale

<table>
<thead>
<tr>
<th></th>
<th>Phase 1 System</th>
<th>Past Systems</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Size</strong></td>
<td>50 billion rows</td>
<td>1-2 billion rows</td>
</tr>
<tr>
<td><strong>Performance</strong></td>
<td>500K in 10 seconds</td>
<td>Hours or days for query completion</td>
</tr>
<tr>
<td><strong>Scalability</strong></td>
<td>Superlinear scalability</td>
<td>Not scalable</td>
</tr>
<tr>
<td><strong>Usability</strong></td>
<td>Single homogeneous data set</td>
<td>Unable to query across data sets</td>
</tr>
<tr>
<td><strong>Timeline</strong></td>
<td>Direct query</td>
<td>Must batch process before query</td>
</tr>
</tbody>
</table>

Phase 1 completed on schedule and under budget
Big Data Initiative Summary

- TRMC is acting upon the KM recommendations from the *Comprehensive Review of T&E Infrastructure*. Strategic Goals:
  - Integrate T&E infrastructure into cohesive Knowledge Management enterprise
  - Modernize T&E practices & processes to leverage Big Data analytics techniques
  - Apply Big Data analytics tools & techniques to the T&E mission space

- TRMC-funded proofs of concept will deliver proven capabilities
  - Enable Big Data analytics during JSF T&E
  - Improve transfer of knowledge between fielded and next-gen systems
  - Inform T&E investment plan that advises future infrastructure, process, and workforce decision-making

- Improved T&E KM will help achieve better acquisition outcomes and reduce costs
  - Identify & Diagnose problems sooner and continuously
  - Inform acquisition decisions through larger knowledge base
  - Achieve T&E infrastructure efficiencies
Questions?

Ryan Norman
TRMC Initiative Lead for Big Data & Knowledge Management
PM, TENA SDA
ryan.t.norman.civ@mail.mil