Deep Data Analytics in Support of Acquisitions and Tradespace Analysis

Engineered Resilient Systems Track
NDIA Systems and Mission Engineering Conference 2018
October 24, 2018

David Stuart
Associate Technical Director
Information Technology Laboratory
Topics

• Current State of Tradespace Analysis
• New thinking
• Technology
• Data analytics ecosystem and processes
• A.I.
• Machine Assisted Tradespace Analysis
Current State

Human Interface (SMART GUIs)

- Tradespace Analysis
  - Dimensionality Reduction
  - Feature Selection
  - Predictive Analysis
  - Descriptive Analysis

- Computational Workflow
  - Simulation Modeling
  - Numerical Modeling
  - Analytical Modeling
  - Optimization Modeling

- Data Integration and Preprocessing

- Data

- Infrastructure

- Meta - Digital Thread / Twin

Maturity

- High
- Low

Current State
New Thinking (10x)

Improve decision making through the integration of advance computing into the decision-making process. Humans perform higher-level strategic thinking, while machines conduct lower-level decision making.

Change Drivers

Today

- Marginalize
  - Analysis is severely restricted
- Limited
  - Analytics do not scale to large problems
- Analysis is very swallow
- Brittle
  - Single point solutions

- Growth in information requires machines to take a more activity participate in decision making
- Humans will conduct high-level decisioning ...machines work to make lower-level decisioning
- Data sizes will overwhelm decision-makers and complicate the decision making process
- Deep Analytics - breadth and depth of the analysis help service insights from all types of data
- Capable of operating on data sets at the petabyte scale
- Prioritize important points for analysis by humans
- decisions will be subdivided into levels machine-level and human-level

Tomorrow

- Go Faster
  - Host analysis on HPC
- Think Deeper
  - Methods scale to address large complex problem spaces
- Inclusion of a breadth of information
- Data and knowledge are integrated
- Be Resilient
  - Identify a set of alternatives as opposed to a single solution
### Competencies

<table>
<thead>
<tr>
<th><strong>Hardware</strong></th>
<th><strong>Software</strong></th>
<th><strong>Data</strong></th>
<th><strong>IT</strong></th>
<th><strong>Policy</strong></th>
<th><strong>Expertise</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Machines that can address our largest problems</strong></td>
<td><strong>Leverages open source capabilities</strong></td>
<td><strong>Terabytes collected - Unorganized and inaccessible</strong></td>
<td><strong>Facilities</strong></td>
<td><strong>Integrate necessary policies</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Blended computing architectures:</strong></td>
<td><strong>Python , R, and C/C++ (when needed)</strong></td>
<td><strong>Streamline data wrangling</strong></td>
<td><strong>Storage (hot, warm, and cold)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Numerical - distributed, high speed interconnects</strong></td>
<td><strong>Anaconda - package management</strong></td>
<td><strong>Minimize the movement of data</strong></td>
<td><strong>Networks (10G+)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Data - shared memory, fast I/O</strong></td>
<td><strong>Jupyter Notebooks</strong></td>
<td><strong>Leverage database technologies</strong></td>
<td><strong>VDI</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Computational and data scientists assist in problem step up, execution, and visualization</strong></td>
<td><strong>Spark, Galaxy, Dakota</strong></td>
<td><strong>SQL and noSQL</strong></td>
<td><strong>Administration of machines</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Data Analytics</strong></td>
<td><strong>Scikit Learn (machine learning)</strong></td>
<td></td>
<td><strong>Security - (monitoring, patching, etc.)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Competencies</strong></td>
<td><strong>TensorFlow (deep learning)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data Analytics Ecosystem

**Mass Storage Services**
- ERDC’s MSAS System (Gold)
- 1PB SAMFS filesystem
- 25PB current on tape (80 PB capacity)
- 3 Oracle T54 servers running Solaris 11
- Data transfers are equally distributed among the 3 T54 servers

**Data Analytics**
- Anaconda Enterprise
- Spark
- Machine Learning
- Deep Learning

**General Computing Services**
- High Speed Network (100G)
- Virtualization
- Persistent Services
- Data Management
- Interactive Access

**Computational Science**
- 126,192 compute cores
- 544 Knights Landing nodes
- 32 GPGPUs
- 437 terabytes of memory
- Adding Urika software stack

**Models**
- Dakota
  - (optimizer)

**Galaxy**
- (workflow)

**Jupyter Notebooks**
- Web connection

**Decision Portal**
- Web or VDI connection

**Decision Makers**

**Data Scientists**

**DGX-1**

**Computational Science & IT**

**Data Science & Decision Making**
Data Science Workflow

Data Science Framework

Data Preparation
- Acquire data
- Reformat & clean data

Software Preparation
- Configure Models & Simulations
- Configure Dashboard
- Configure Tools
- Configure Workflow

Analysis
- Execute Workflow
- Inspect Outputs

Reflection
- Compare Alternatives
- Hold Meetings

Dissemination
- Deploy Dashboard
- Write Reports

Tools
- jupyter
- ANACONDA
- GALAXY

- Data Insights
- Data Science Framework
- Tools
- Analysis
- Reflection
- Dissemination

~ 3 months

US Army Corps of Engineers • Engineer Research and Development Center
Distribution A: Approved for public release
Current Tradespace Workflow

- Heavy Visualization
- Manual process
- Millions of designs considered, but only a few in detail
With data…

Without data…
With Data - Virtual Sensors

Historical operational sensor data can be used to study how many sensors are needed and where they need to be placed.

If a sensor can be inferred from other sensors with a high degree of accuracy then instead of fielding a physical sensor, a “virtual sensor” model, developed on DSRC-HPC, can be used.

The minimum virtual sensor cover is the minimum set of physical sensors necessary to infer ALL sensors with some required level of accuracy.

A **minimum virtual sensor cover** for FVL would save space, weight and power, extending range and lifting capacity and **save hundreds of millions of dollars** in up front manufacture and life cycle maintenance cost.
Without data – AlphaGo Zero

- Learn from scratch
- No historical data
- Data generated from unsupervised training
- 20 days of training to beat world champion

Complexity
Chess: $10^{120}$
Go: $10^{174}$
Machine Assisted Tradespace Analysis - Needs

- Full definition of the problem
- Need win condition and rules
  - Capabilities
  - Constraints
- Functional Framework (Driver)
- No beginning tradespace
Ontologies

- Machine Assisted Driver
- Provides Structure
- Defines Semantics
  - User understanding
  - Machine understanding
- Defines constraints
- Drive digital twin
Map Ontologies to models

- Attributes within the ontology are mapped to inputs of the available models
- Performance metrics mapped to outputs
Machine Assisted tradespace analysis
How does it work

First Iteration
- Low fidelity
- Millions of designs
- Clusters of high performance designs
- Select 1 or more clusters

Win condition – desired capabilities
Machine Assisted tradespace analysis
How does it work

Second Iteration
• Moderate fidelity
• 10,000’s of designs
• Clusters of high performance designs
• Select clusters within clusters

Win condition – desired capabilities
Machine Assisted tradespace analysis
How does it work

Nth Iteration
- Increasingly higher fidelity
- Decreasing number of designs
- Clusters of high performance designs
- Select clusters within clusters

Win condition – desired capabilities

Final output
- Set(s) of designs
- Complex constraint sets
Contact

David Stuart
US Army Engineer Research and Development Center
Information Technology Laboratory
david.c.stuart@erdc.dren.mil
601-634-3174