Abstract

Researchers with the Engineered Resilient Systems (ERS) program are engaged in multiple efforts to effectively utilize large data sets collected from DoD platforms to apprise agencies of system performance, improve reliability and availability, and inform future requirements. The foundational technology for this work is a High Performance Computing (HPC)-based infrastructure that supports large data management – a data lake ecosystem. A *data lake* is a repository of related data that is maintained in its original format. Any transformations performed on this data result in a new *pool* of data on which analytics can be executed. The original and derived forms of data, together with the supporting tools and technologies, comprise a data lake ecosystem. This ecosystem supports high performance, parallel analysis of large data sets, and facilitates data provenance and access controls. Large-scale data analytics projects include maintenance data analysis for reliability assessment, and model development for impacting future design. For example, researchers are currently investigating the ability to infer the output of a "virtual sensor" from an actual sensor that is in close proximity. This capability has two primary use cases: first, in existing vehicles with standard sensor packages, one sensor could detect when another sensor is malfunctioning, increasing safety and facilitating improved maintenance. Second, test data from prototypes could be used for determining the minimum number and optimum placement of sensors, decreasing cost and operational weight. Other efforts in this field include demonstrating cross-service applicability of machine learning models to maintenance data for natural language processing and prediction capabilities, and using large data sets to create surrogate models to replace computationally intense, long-running codes. The ability to effectively analyze complete historic data sets also enables an accurate verification of algorithms that were previously developed on information based on much smaller samples of data.

Techniques in Large-scale Data Analytics for Engineers

R. Cody Salter, D.Sc.

22nd Annual Systems and Mission Engineering Conference

23 October 2019

US Army Corps of Engineers

ERDC
ENGINEER RESEARCH & DEVELOPMENT CENTER

Presentation Outline

1. Introduction
2. Technical Domains
3. HPDA Applications
4. Data Lake Ecosystem
5. HPDA Infrastructure
6. Conclusions

Introduction

- As modern weapon systems continue to advance technologically, the amount of data they produce increases at a rapid rate.
- If stored, maintained, and utilized properly, weapon system data can provide useful insights on system performance and reliability, thus impacting availability, soldier safety, and lifecycle costs.
- Because of these benefits, the Engineered Resilient Systems (ERS) program is engaged in multiple data analytics efforts.
  - ERS hopes to provide meaningful insights to DoD platforms while also developing the infrastructure needed to support future data analytics research.

Technical Domains and Applications

- **Overall, ERS data analytics projects have focused in four technical domains:**
  - **Large-Scale Data Management**
  - **Deep Learning and Machine Learning**
  - **Natural Language Processing**
  - **Automated Data Labeling**
- **The application of these technical domains to DoD problems has resulted in many meaningful capabilities:**
  - **Automated rotorcraft logbook labeling**
  - **Rotorcraft flight regime recognition**
  - **Large-scale ground vehicle data analytics**

Automated Rotorcraft Logbook Labeling (1)

Problem: Vehicle operational reliability data is not directly available from maintenance logbook entries

Solution:

- Use Natural Language Processing (NLP) to infer engineering reliability data from maintenance logbook data

Impact:

- Enable 100% of logbook data to be utilized for engineering reliability assessment
- Significantly reduce time for assessment

Established 92% accuracy of reliability data generation from logbook entries

Maintenance Events

Logbook Reports

Analyst

ONLY
10% scored

Automated Rotorcraft Logbook Labeling (2)

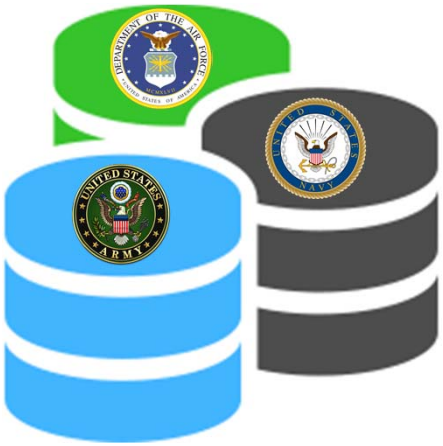
Problem: Army, Navy and Air Force use different terminology for maintenance on the same rotorcraft platform

Solution:

- Use deep learning to discover related terms based on semantics

Impact:

- Drastically increase data availability for analysis and model development
- Share models across services
- Step toward standardizing vocabulary



Logbook Reports

Word2Vec

Word2Vec uses deep learning to derive the underlying meaning of text and to produce numeric models for machine learning algorithms

Joint Corpus

Rotorcraft Flight Regime Recognition

Problem: Current flight regime recognition algorithms were developed on small data samples and have unknown accuracy

Solution:

- Use historic sensor data to develop a machine learning model for recognizing flight regimes

Impact:

- Perform data-drive validation of currently used flight regime software
- Enable the accurate assessment of aircraft fatigue and part life

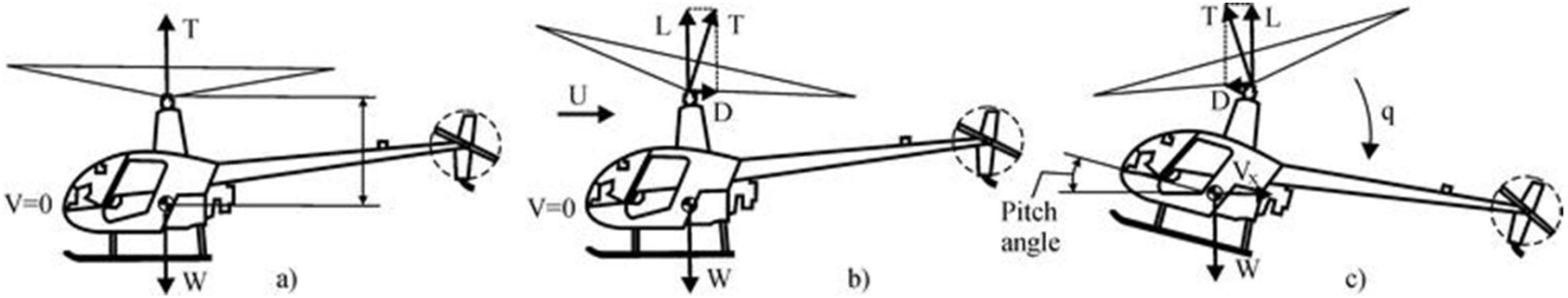


Image from https://www.intechopen.com/books/flight-physics-models-techniques-and-technologies/helicopter-flight-physics

Large-Scale Ground Vehicle Data Analytics

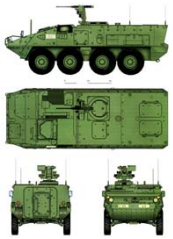
Problem: DoD lacks the ability to fully analyze the large data generated during vehicle operations

Solution:

- Develop a reproducible process for ingesting and analyzing very large ground vehicle data sets

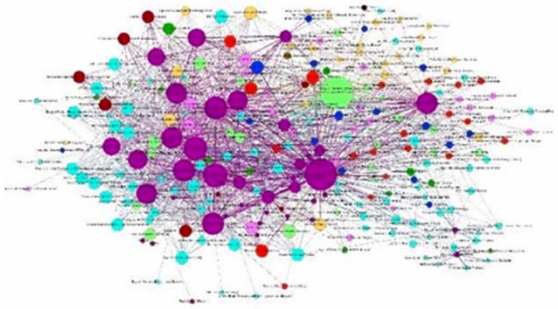
Impact:

- Reduce maintenance costs
- Increase vehicle/component useful life
- Improve analysis and reporting
- Automate ingestion and rollup reports
- Improve reliability, availability, and maintainability (RAM)



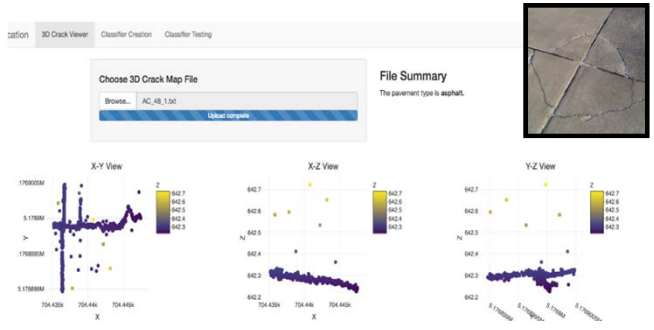
Applications of High Performance Data Analytics

Vehicle Maintenance



Cyber Analytics

Image Analysis



Airfield Quality Assessment

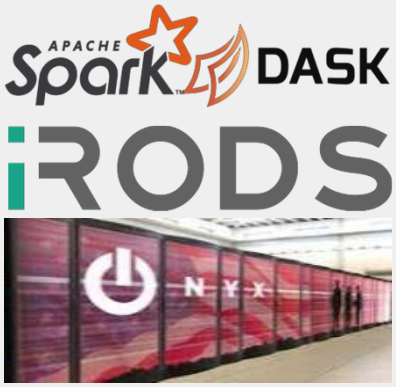
Waterways Data Mining

Tradespace Analysis

High Performance Data Analytics Infrastructure

HPC Infrastructure

Implement supporting frameworks for Data Lake Ecosystem and user interfaces

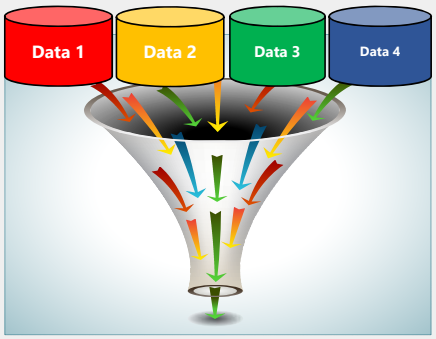


Application Workflow Management

Procure, install, and ensure functionality of data analytics and data management software

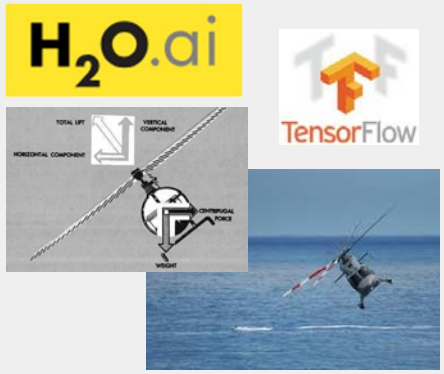
Data Management

Procure and integrate data sets into the Data Lake Ecosystem for analytics use



Data Analytics

Provide data science and AI/ML expertise for analytics projects



Conclusions

- **The Engineered Resilient Systems program is engaged in multiple data analytics efforts in hopes of bringing data analytic and machine learning capabilities to programs and platforms across the Department of Defense.**
- **ERS funding has supported technical development in four primary data analytics domains, reaching across several application areas to include maintenance logbook labeling, algorithm verification, and ground vehicle data analysis.**
- **Overall, ERS supports research focused on future innovation in the field of high performance data analytics.**

ERS
ENGINEERED RESILIENT SYSTEMS
DEPARTMENT OF DEFENSE

ERDC
ENGINEER RESEARCH & DEVELOPMENT CENTER

Questions?

R. Cody Salter, D.Sc.

Research Mechanical Engineer
Institute for Systems Engineering Research (ISER)
Information Technology Laboratory (ITL)
Engineer Research and Development Center (ERDC)

Richard.C.Salter@erdc.dren.mil
(601) 619-5152

DISCOVER • DEVELOP • DELIVER

new ways to make the world safer and better

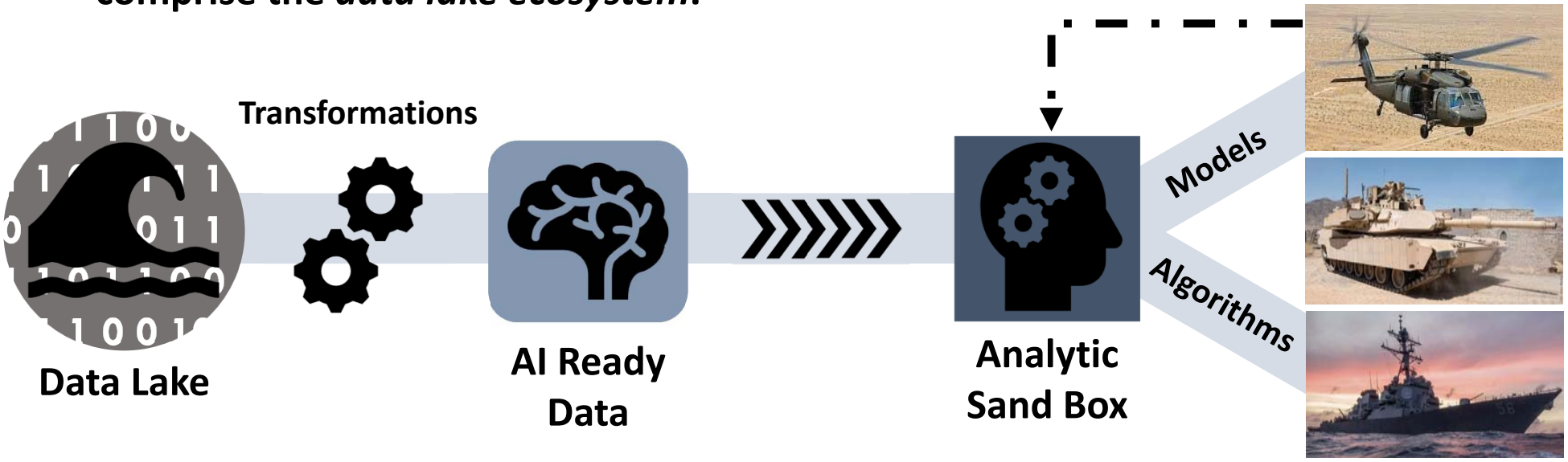
What is a Data Lake?

- **The foundational technology for this work is a High Performance Computing (HPC)-based infrastructure that supports large-scale data management known as the *data lake ecosystem*.**
- **What is a *data lake?***
  - **According to Amazon Web Services, a data lake is "a centralized repository that allows you to store all your structured and unstructured data at any scale."**
  - **The IBM Corporation utilizes a similar definition: "data lakes are next-generation hybrid data management solutions…their highly scalable environment can support extremely large data volumes and accept data in its native format from a wide variety of data sources"**
- ***Data lake* – an immutable repository of related data that is maintained in its original format.**

Amazon Web Services. (2019). What is a data lake? Retrieved June 26, 2019, from https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/
IBM Corporation. (2018). Build a better data lake. Retrieved June 26, 2019, from https://www.ibm.com/analytics/data-lake

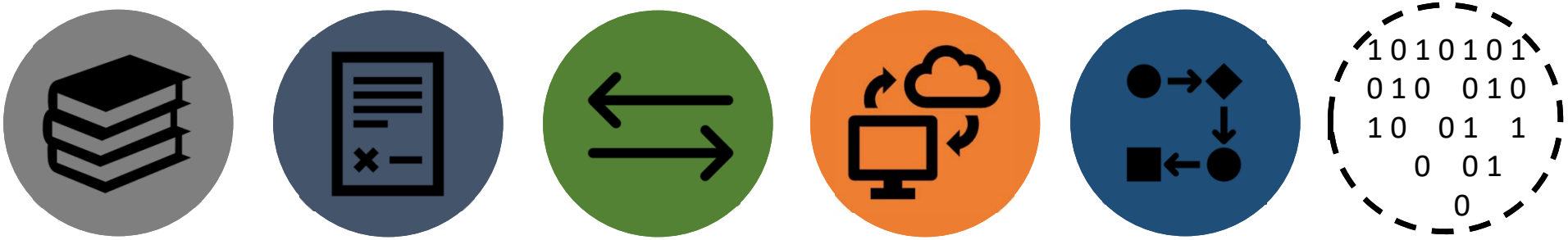
Data Lake Ecosystem

- As original data is transformed, derivative data sets are created for future analytics.
- Original and derivative data, combined with supporting tools and technologies, comprise the *data lake ecosystem*.



Data Lake Ecosystem Workflow

- Data transfer between government organizations is often a difficult and time-consuming process requiring the participation of multiple parties.
- As a result, our team set out to document a data transfer workflow in hopes of improving efficiency, educating collaborators, and educating team members.



Educate & Interact | Access & Agreements | Transfer & Ingest | Store & Update | Transform & Analyze | End of Life