

Raytheon

JPSS Cloud Data Processing Future and Machine Learning for the Weather Value Chain



Scott Kern
Emily Greene
Shawn Miller

Copyright © 2018, Raytheon Company. All rights reserved.

This document does not contain technology or technical data controlled under either the U.S. International Traffic in Arms Regulations or the U.S. Export Administration Regulations.

Agenda

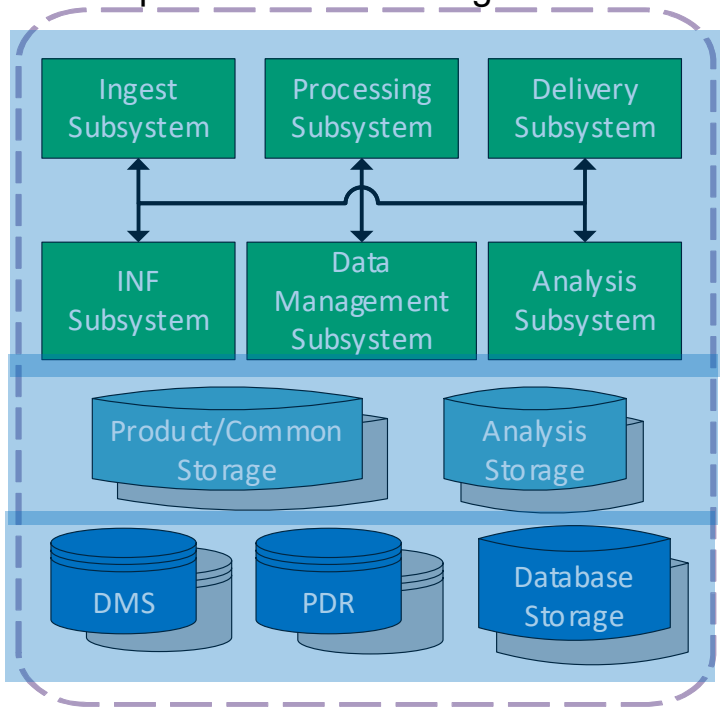
- IDPS in the Cloud for NESDIS
- Optimization Plans
- Cloud enabled Opportunities
- Machine Learning for Operational Weather

This document does not contain technology or technical data controlled under either the U.S. International Traffic in Arms Regulations or the U.S. Export Administration Regulations.

Initial Implementation – Phase 1

- Transition to Operations in Cloud must occur NLT EOY 2020 (Lenovo HW waiver expiration)
- NOAA direction to migrate current operational baseline to Cloud with minimal baseline changes
  - Only changes to baseline that are explicitly necessary to operate in the cloud
    - Moving primary IDP DB from Oracle to PostgreSQL to save Oracle licensing costs
- HOT backup of primary Operations IDP
  - Monthly Security Patching requires transition to backup IDP
  - 3rd IDP necessary to accommodate monthly patches and baseline upgrades while maintaining resiliency to failures
- Primary change is new Common Environment :
  - Route data to multiple IDPS systems from a single on-prem data source
  - Management of security solutions
    - On-prem IDPS is a “user” of security solutions from the C3S segment
- Leveraging DevOps Tools/Processes:
  - Environments 100% managed using Infrastructure-as-Code (Packer, Terraform, Chef)
  - Faster/Frequent algorithm releases to PRO subsystem decreases Research-to-OPS (R2O) cycle
- ~60 EC2 VMs and 500 TB EBS storage per Ops-capable IDPS

Operational IDP Configuration



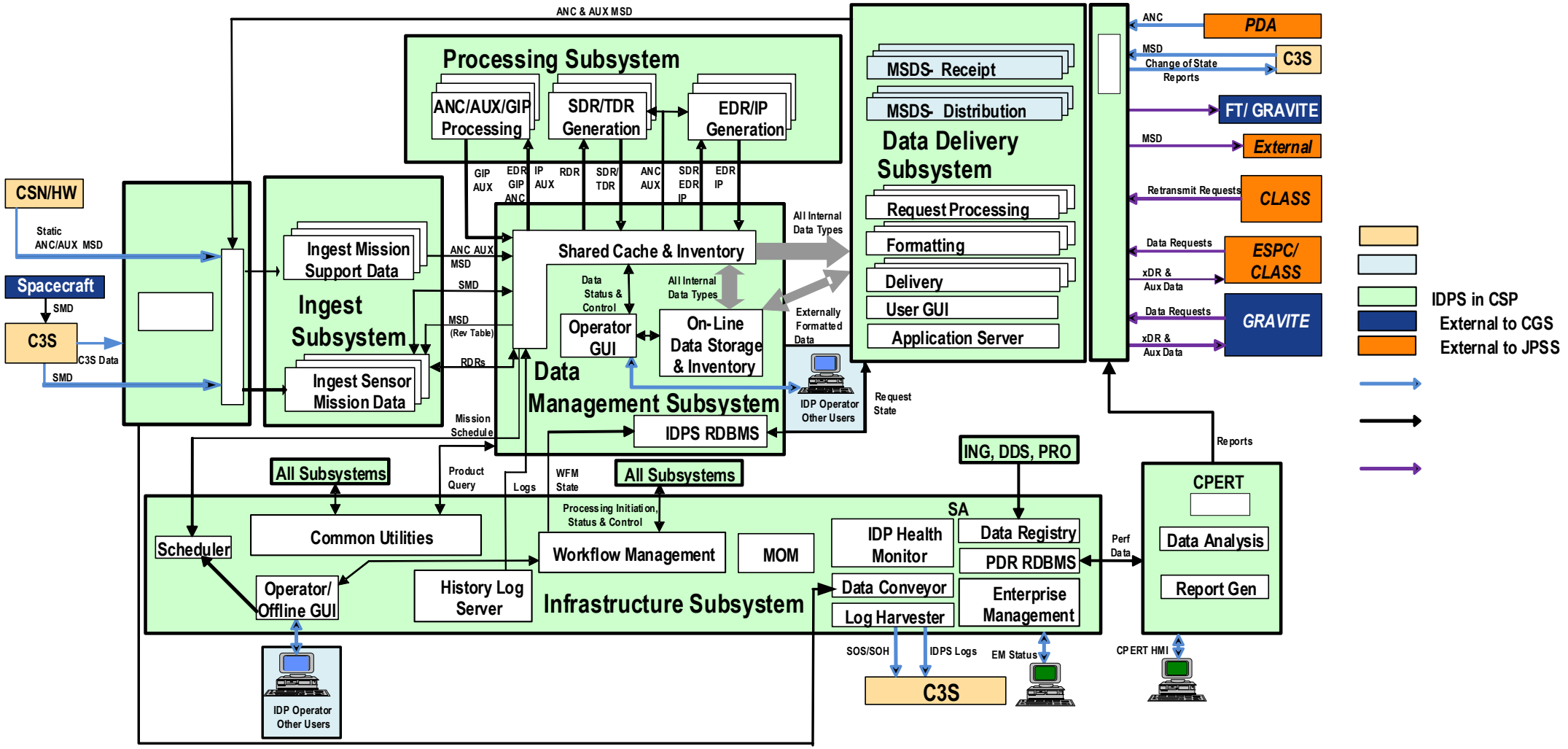
Database Layer (EC2 and EBS)
- Oracle Dataguard installed to EC2
  - Backup DB instance
- EBS storage attached to EC2

- DMS: Data Management
- PDR: Performance Data Repo

This document does not contain technology or technical data controlled under either the U.S. International Traffic in Arms Regulations or the U.S. Export Administration Regulations.

IDPS in the Cloud Architecture Overview

CGS-018



This document does not contain technology or technical data controlled under either the U.S. International Traffic in Arms Regulations or the U.S. Export Administration Regulations.

AWS Services In Use For Initial Implementation

AWS Service	Purpose
EC2/EBS	• Processing VMs, significant tuning in Task Order prototypes to define the configuration • SIGNIFICANT volume of EBS storage required for GPFS, 250 TB writable data perf IDP. GPFS installed to EC2 however needs full replication for resiliency to a single EC2 failure, 500TB total storage.
PostgreSQL Relational Database Service (RDS)	• Hosts primary data management database (DMS) • Deployed in multi-AZ configuration for redundancy, but other IDPS components are single-AZ
Simple Storage Service (S3)	• SMD and MSD archive hosted in cloud DPC • Factory use for testdata and other large datastores • Storage for Artifactory COTS in deployment pipeline • Drop-box for algorithm changes in DP-AE
Simple Notification Service (SNS) Simple Queue Service (SQS)	• Used by new Mission Data Distribution function to deliver one data source to many IDPs • New SMD/MSD product arrives in S3, SNS sends message to a SQS queue assigned to each IDP, guarantees each IDP receives all SMD/MSD even when not active
CloudWatch	• Monitoring and aggregation of cloud logs (from CloudTrail, VPC Flow Logs and CloudWatch agents) • Delivers security relevant events to on-prem Qradar
CloudTrail	• Logs API calls to AWS services
VPC Flow Logs	• Logs traffic in/out of each VPC
Direct Connect Gateway (DXG) Transit Gateway (TXG)	• DXG enables direct connect at NWAVE • TXG routes data from DXG to each VPC

This document does not contain technology or technical data controlled under either the U.S. International Traffic in Arms Regulations or the U.S. Export Administration Regulations.

Optimization – Phase 2

- Optimization Phase Updates the IDPS cloud design to take better advantage of cloud capabilities
- Provides significant cost savings over initial-implementation
  - Savings for Infrastructure, COTS, O&M
- Implements a better foundation for science/forecast product driven changes during Modernization Phase

Optimization	Description

Transition to Highly Available (HA) IDPS	• Deploy single HA IDPS spanning 2 Availability Zones
	  • Subsystems deployed across AZs in auto-scaling groups
	  • "Live" security patching on dynamic instances to eliminate OPS/Non-OPS transitions for monthly security patching

Dynamic Allocation of Processing Capacity	• Elastic processing capacity to dynamically respond to changing throughput needs in responding to anomalies
Complete migration of all databases to PDR	• COTS licensing savings • Reduces DBA support needs and security patching overhead

Modernize IDPS Storage Layer	• Product storage moved from GPFS to cloud-native blob storage (AWS S3)
	  • Significant cost savings
	  • Initial prototyping shows satisfactory performance with minimal code modifications
	• Common storage migrates to cloud-native shared file system (AWS Elastic File Service EFS)
	  • Provides HA without overhead required to manage large replicated storage cluster

Utilize Clustered Messaging Service	• Develop HA messaging system or utilize "Messaging-as-a-Service from AWS (Amazon MQ)
Utilize Cloud-Native Monitoring and Alerting	• Initial-Implementation using legacy design of monitoring agents deployed on IDPS VMs delivering messages to operations.

BLUE: Denotes successfully prototyped and demonstrated capability

This document does not contain technology or technical data controlled under either the U.S. International Traffic in Arms Regulations or the U.S. Export Administration Regulations.

Modernization – Phase 3

- The modernization phase leverages IDPS' proven data production platform
  - Provide an expanded number of enterprise data products
  - Decrease algorithm process overhead accelerating R2O cycle
- Data Delivery capability to expanded user base while minimizing data egress costs
  - Prioritize Real-time products critical to NWP delivered with IDPS' proven low-latency and stability
  - Non-Real-time critical products have packaging and delivery processing

Optimization	Description

Modernize Processing Subsystem using Containerized Algorithms	• Science teams will directly develop algorithms and include dependencies in versioned containers
	• Run multiple algorithm versions in parallel, dependencies reside in container
	• Enterprise data product generation
	• Real-time Processing: Operational algorithms generating products
	• Off-line Processing: "Algorithm Sandbox" Evaluate updates to algorithms
	  • Executed during "back-orbits", spot-instances or serverless
	  • Eliminates need for full IDPS dedicated for dedicated I&T and provides faster R2O cycles

Modernize Data Delivery via Cloud-based Content Delivery Network	• Data products delivered to single cloud location (S3)
	  • Eliminate delivery of products through C3S facility to Mission Partners
	• Real-Time Delivery: Products delivered to S3 location
	  • NWP products delivered in directly ingestible format (HDF, BUFR, NetCDR, etc)
	  • Consumers who need real-time products will receive notification of new products and API to pull the data directly down to their system (S3 => SNS => SQS pipeline)
	• Off-Line Delivery:
	  • Non-Real-Time consumers will be able to request aggregation and/or packaging of products which will create a new product in S3 and notification delivered to consumer

"Lights Out" IDPS decreases reliance on dedicated operations staff	• IDPS is highly stable system requiring almost no human interaction to function
	  • Decreases reliance on 24x7 dedicated operators
	• Remove Java based GUIs and replace with simplified web GUI with APIs to drive IDP functions
	  • Significantly improves security posture

This document does not contain technology or technical data controlled under either the U.S. International Traffic in Arms Regulations or the U.S. Export Administration Regulations.

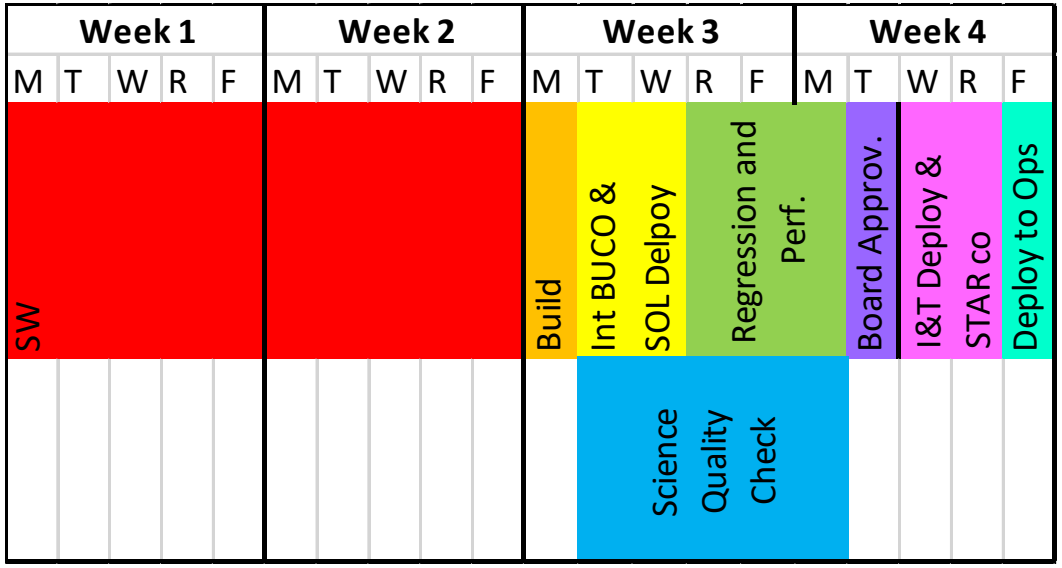
Cloud Enabled Opportunities

- Data Delivery via Cloud Storage
  - Data products delivered to single cloud location (S3)
    - Eliminate delivery of products through C3S facility to Mission Partners
  - Real-Time Delivery: Products delivered to S3 location with no aggregation/packaging
    - Significant simplification of delivery subsystem
    - Consumers who need real-time products will receive notification of new products and API to pull the data directly down to their system
    - S3 => SNS => SQS pipeline
  - Off-Line Delivery:
    - Non-Real-Time consumers will be able to request aggregation and/or packaging of products which will create a new product in S3 and notification delivered to consumer

This document does not contain technology or technical data controlled under either the U.S. International Traffic in Arms Regulations or the U.S. Export Administration Regulations.

Cloud Enabled Opportunities

- Containerize Processing Algorithms
  - Science teams will directly develop algorithms and include dependencies in versioned containers
  - Run multiple algorithm versions in parallel, dependencies reside in container
  - Real-time Processing: Operational algorithms generating products
  - Off-line Processing: Evaluate updates to algorithms, executed during “back-orbits” or serverless
    - Less capacity required than real-time
    - Eliminates need for full IDPS dedicated for dedicated I&T and provides faster Science-to-OPS cycles
- Rapid Science Container updates
  - 4 weeks cycles to reach Operations, 2 cycles in development at all times
  - SW with 2 week sprints
  - Extra time for Performance check and Science Quality evaluation



This document does not contain technology or technical data controlled under either the U.S. International Traffic in Arms Regulations or the U.S. Export Administration Regulations.

General Application of Machine Learning

SYSTEMS ENGINEERING

Where in Processing Chain should AI/ML be Applied?

- Benefit vs. cost
- Data availability – quality & quantity
- Impact on overall system



SOFTWARE/ ALGORITHMS

Interface, library or tool which allows developers to more easily and quickly build machine learning models, using pre-built optimized components

DEVELOP/TRAIN MACHINE LEARNING ALGORITHMS

- Select algorithm, language, framework
- Establish training environment
- Train neural nets, evaluate results, iterate

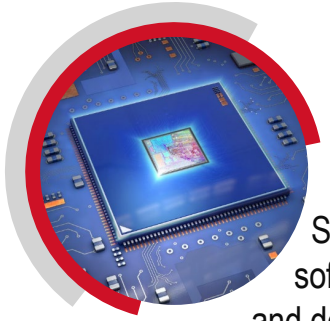


DATA

Information that contains a signal about the event you care about. AI/ML data can be in the form of images, text or video. Training data: data that we present AI/ML algorithms to train them; Test data: data that we use to evaluate accuracy of AI/ML algorithms

IMPLEMENT ALGORITHMS

- Select processing approach
- Integrate with existing system



HARDWARE

Specialized hardware to accelerate software which runs machine learning and deep learning applications

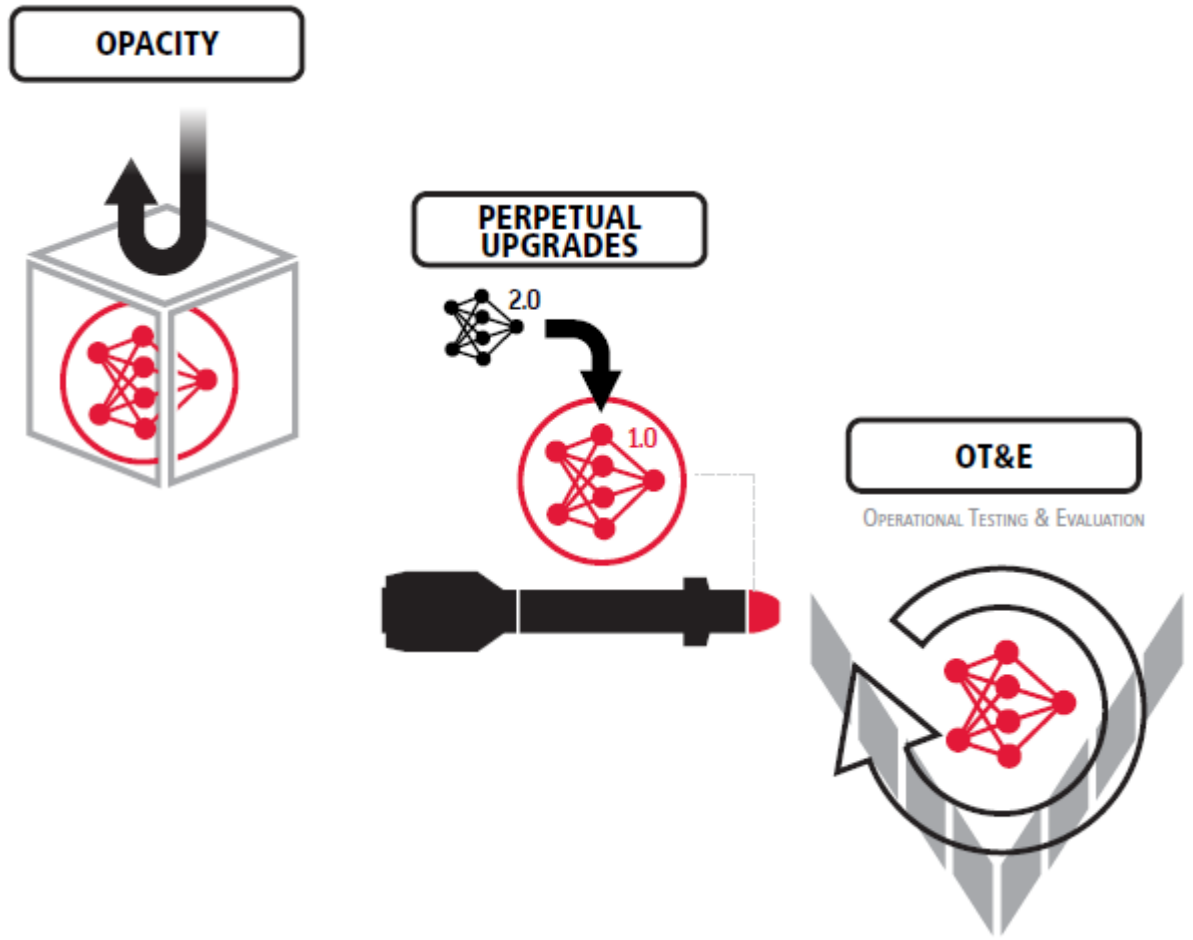
EVALUATE SYSTEM

- Select evaluation data
- Perform Verification & Validation, operational testing

This document does not contain technology or technical data controlled under either the U.S. International Traffic in Arms Regulations or the U.S. Export Administration Regulations.

Challenges to Machine Learning

- Opacity (i.e., “the black box”):
  - Explaining why ML got an answer is just as important as getting the answer
- Perpetual Upgrades:
  - Evolving requirements can cause models to have a life cycle of mere seconds; key discriminator is automation
- Operational Test and Evaluation:
  - ML models are inherently complex, non-deterministic systems; potential exists for unanticipated emergent behavior, indeterminate test results, Black Swan events (another fertile ground for innovative solutions)



This document does not contain technology or technical data controlled under either the U.S. International Traffic in Arms Regulations or the U.S. Export Administration Regulations.

Improving Machine Learning Itself

- Strategies for Rapid Prototyping Machine Learning
  - Difficulty in pattern recognition is the sheer volume of source data that must be analyzed – time required to acquire, label, and train the model
  - We have developed a number of tools and approaches to maximize training efficiency and mitigate effects of limited training exemplars, bad labels, and noisy data
  - Example 1: integration of Generative Adversarial Networks (GANs) into the training process (joint probability between inputs and outputs); shown in one study to reduce training iterations by factor of 10
  - Example 2: Pseudo-Labels (labels that are created automatically for unlabeled data using a partially trained network) can significantly improve classification accuracy without changing network architecture, based on theory known as Entropy Regularization

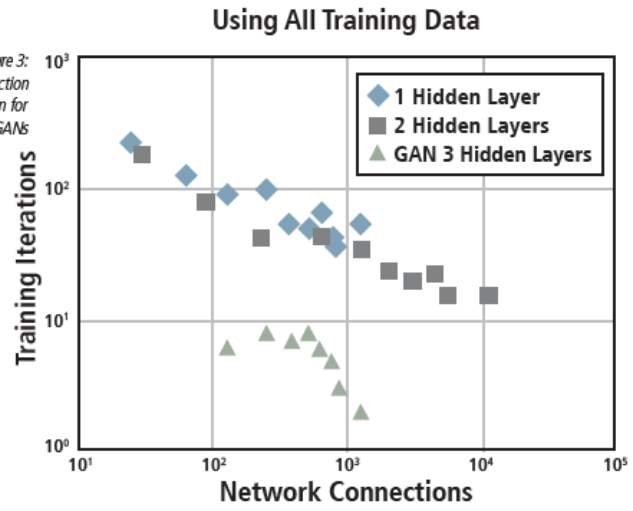


Figure 3: Significant reduction in training burden for GANs versus non-GANs

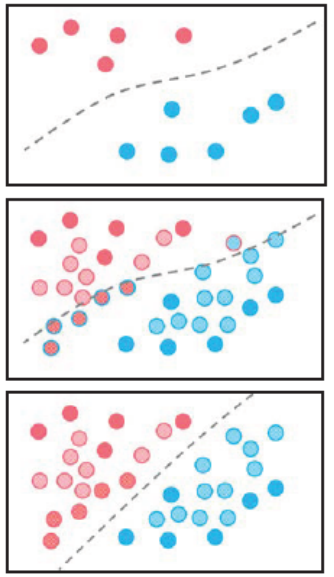
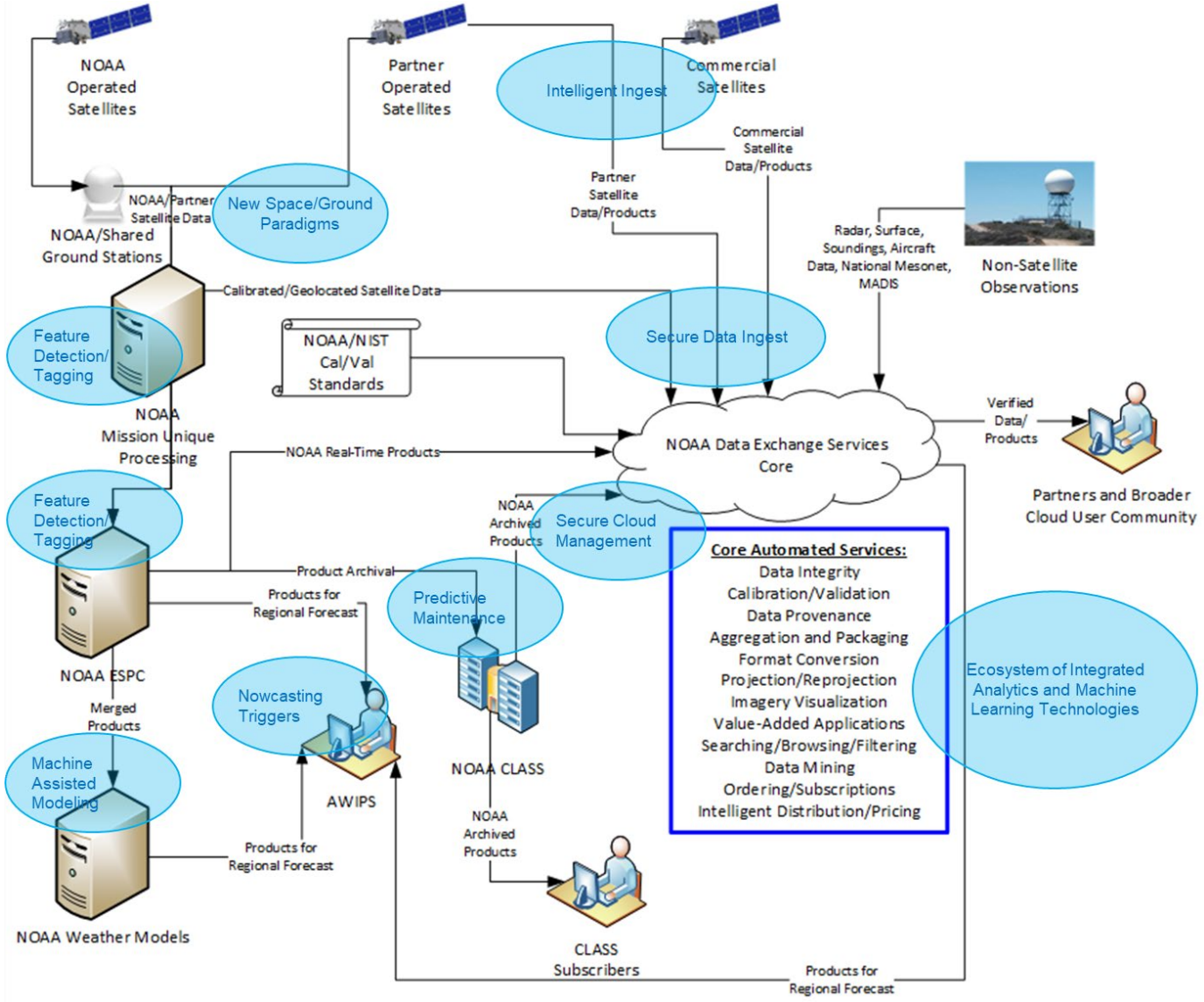


Figure 4: TOP: Sparse labeled data has ambiguous class boundaries, MIDDLE: Unlabeled data with pseudo-labels (outlines) added to dataset, BOTTOM: Re-training with pseudo-labels corrects decision boundaries based on data population density.

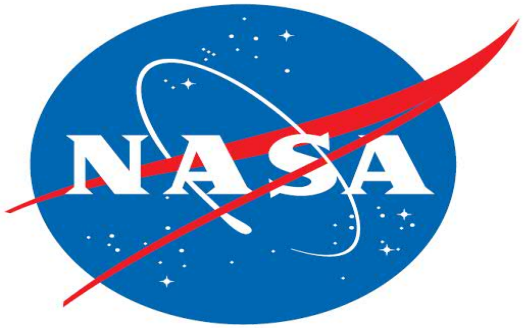
This document does not contain technology or technical data controlled under either the U.S. International Traffic in Arms Regulations or the U.S. Export Administration Regulations.

Analytics, ML, and the Weather Value Chain



Impact Area	Forecast Benefit
Accelerated Data Ingest	• Commercial data available for forecast models • Reduced data latency
Identify high-value data for forecast models during sensor processing	• Improve initial conditions for model runs • Improved Satellite Data Assimilation
Training models for weather anomalies	• Increase speed and accuracy of extreme weather forecasts • Provide forecaster tools to train regional datasets
Federate Weather data to process disparate data sets in the cloud	• Faster R2O2R due to removal of data stovepipes • “multi-int” data fusing across commercial/govt data types
Cloud technologies (massively parallel systems, GPUs, cloud computing)	• ML for all classes of users, with elasticity for cost control • Better collaboration across weather enterprise community

This document does not contain technology or technical data controlled under either the U.S. International Traffic in Arms Regulations or the U.S. Export Administration Regulations.



Raytheon



This document does not contain technology or technical data controlled under either the U.S. International Traffic in Arms Regulations or the U.S. Export Administration Regulations.