Outline of Talk

• Opportunities
• What’s it Going to Take
• Analytic Centers as a Framework
• Concept of Operations
• Some Experiments
• Strategy
Technology Opportunities

• Generating data and numeric output
  – Constellations of SmallSats with in situ, airborne platforms
  – Closer interaction with physical and statistical models
  – Autonomy

• Computing Capabilities
  – Public Cloud Computing
  – Quantum Computing
  – Cognitive Computing

• Exploiting Data
  – Machine Learning
  – Workflow tools, such as Notebooks
  – Integration of data into a unified picture of a natural phenomenon or physical process
  – Virtual Reality
Machine Learning Hopes

• A more robust **understanding** of a natural phenomenon or physical process
  – Relying on conclusions based on 3,000 examples, not 3
  – Digest massive volumes of observational data
  – High volume output from high-resolution ES models
  – Conventional techniques drown ES researchers

• **Prediction** of future state for complex systems from observational data without a thorough understanding of complex or poorly understood underlying physics

• **Coordination** of elements of a sensor web to target transient and transitional phenomena
  – Leverage emergence of SmallSats
  – Autonomy
  – Detection of interesting features and re-tasking to observe them

• Fast, comparative **OSSEs**
  – What is the relative value of science for different Observing Systems

• **Assimilation** models which do better at washing out the experimental error while preserving features and physical anomalies

• A way to **clean up observational data** to make it easier to analyze
Lessons Learned (so far)

• ML and advanced tools are not trusted by Science Community
  – Not well understood, including which ones to use for what
  – Difficult to use
  – Don’t know who to trust

• Current crop of tools are hard to use

• Data is hard to coerce into an analyzable format
  – Finding usable data is hard (current catalogs offer limited help)
  – Large volumes of data are hard to process
  – Diverse formats, uncertainties, sources are hard to integrate

• Non-Earth Science Community is far ahead in using them
  – Financial Markets
  – Computer Security (and other security)
  – Commercial Space
What’s it Going to Take?

• **Community Acceptance** in NASA Earth Science (ES)
  – NASA Earth Scientists must understand the processes & algorithms
    • “In the field of observation, fortune favors the prepared mind.” – Pasteur
    • Collaboration among Earth Scientists, Computer Scientists and Technologists
  – Demonstration of value to the science communities
    • New Science
    • Faster Time to Science, retaining quality and understanding
    • Results backed up by confirmation with legacy techniques
    • Uncertainty Quantification
  – Engagement with NASA ES by experienced ML practitioners

• **Improved Data Usability**
  – Rapid use of the numeric results without time-delays for grooming

• **Easy to use** tools and credible results
  – Need a Framework to integrate tools, data and computation resources
A Way to Focus Technology Development

• An environment for conducting a **Science investigation**
  – Enables the confluence of resources for that investigation
  – Tailored to the individual study topic

• Harmonizes **data, tools and computational resources** to permit the research community to focus on the investigation
  – Reduce the data preparation time to something tolerable
  – Catalog of optional resources (think HomeDepot shopping)
  – Semantic-enabled catalog of resources (think Yelp) with help
  – Relevant publications
  – Provide established training data sets of varying resolution
  – Provide effective project confidentiality, integrity and availability
  – Single sign-on and unified financial tracking
Analytic Center as a Framework

**User**
- Project Definition
- Plan for Investigation

**Data**
- Catalog
- NASA DAAC
- Other US Govt
- Non-US
- Local or non-public

**Project Work Environment**

**Tools**
- Discovery & Catalog
- Work Management
- Data Interfaces
- Analytic Tools
- Modeling
- Collaboration
- Visualization
- Sharing/Publication
- Local/custom

**Computational Infrastructure**
- Computing
  - Capacity
  - Capability
- Storage
- Communications

**Storage**
- Data Containers
- Thematic model
- Metadata/Ontology
- Resulting Products
- Published data
- Provenance

**Computing**
- Local systems
- High End Computing
- Cloud Computing Capability
- Quantum Computing
- Neuromorphic Computing
AC Concept of Operations

• **Principles**
  – Each researcher defines their own process
  – Community accepted tools allow re-use of others’ work
  – Unique tools and data should be easily integrated
  – Collect publication materials as you go, in background

• **Notional Process**
  – PI defines hypothesis and process
  – Select framework to be used and workflow tool
  – Shop for tools, data and compute resources
  – Assemble components and verify integrity
  – Perform analysis, develop models, visualize, etc.
  – Draw conclusions and perform internal reviews
  – Enter publication process
AIST Experiments with Analytic Centers

• Land Use and Land Change
  – NEX (NASA Ames) Rama Neimani

• Physical Oceanography
  – OceanWorks (JPL) Thomas Huang

• Tropical Cyclones and Hurricanes
  – TCIS (JPL) Svetla Hristova-Veleva

• Climate
  – Climate Model Diagnostic Analyzer (CMDA), Seungwon Lee (JPL)
  – Climate Workbench (UAH/MSFC) Manil Maskey, Chris Lynnes (GSFC)

• Communities being discussed for further experiments
  – Biodiversity
  – Hydrology
  – Atmospheric Composition
  – Cryosphere
NASA Earth Exchange (NEX) as an Analytic Center

Land Change/Use Community

- Project Definition
- Plan for Investigation

Data

- Built into website
- Dataset Sources
  - Landsat
  - Sentinel 1A
  - Modis
  - ASTER
  - TRMM
  - AVHRR
- Climate Datasets
- Land cover
- Digital Elevation Map
- STATSGO Soils
- USDA Aerial NAIP
- And others

Tools

- Models
  - Tops
  - Biome-BGC
  - LPJ Dynamic Global Model
- Sandbox for small scale experiments
- Analysis
  - R and python based tools
  - Matlab VIIRS HDF5 swath conversion
- Workflow: Jupyter Notebook

Computational Infrastructure

Storage

- Data Containers
- Thematic model
- Metadata/Ontology
- Resulting Products
- Published data
- Provenance

Computing

- NASA Advanced Supercomputing (NAS)
- Amazon Web Services (AWS) Public Cloud
- Ames Quantum Computer (D-wave)

POC: rama.nemani@nasa.gov
https://nex.nasa.gov/nex
A Strategy for Introduction

- Focus on needs of a particular Science community
  - Data, models and tools they are comfortable with
  - Computational resources adequate to perform their investigation
  - Concept of Operations must support their work, not force a fit
  - Identify needs, lessons learned, community perceptions
  - Are there common processes that could make it easier

- NASA’s Earth Science needs **much** more exposure to Machine Learning
  - Researchers and Managers need to develop confidence in the tools
  - Outreach to the Science community

- Must be an architecture with modularity and interoperability
- Must support security
- Increase access to non-NASA data sources as well as NASA
- AIST will experiment with basic concepts as well as components
BACK UP
AC ConOps Detailed Sample

- PI defines the investigation to be done
  - Types of data, analysis steps, process
- Select workflow tool to manage process
  - Such as Jupyter Notebook
- Shop for data, tools and computing resources needed
  - Parameter search of known data repositories
  - Find more like...
  - Identify data from local or unpublished sources
    - Create metadata to make it usable by others, including peer review
- Integrate local data into local storage – compatibility mode
- Retrieve remote data into local storage
  - Storage method is harmonized with known tools
  - Verifies accuracy of copy compared to repository
  - Perform incremental copy/analysis if storage limited
- Integrate local tools into work flow to interact with data and storage effectively
  - Document the tool for future reference and publication
- Massage data as required (geolocation, gridding, characteristics)
- Step through analytic steps with a limited subset of data using a limited set of processing resources
  - Visualize or analyze results until satisfied with process
- Scale up processing of entire data set
- Review results and draw conclusions
- Review AC publication package
  - Source code, tools, local data, metadata, etc.
- Submit for publication
Earth Data Analytics System (EDAS) as an Analytic Center

**Data**
- MERRA
- ECMWF ERA
- NOAA NCEP CFSR
- JMA JRA
- UA ERA

**Tools**
- Earth System Grid Federation Framework
- Discovery & Catalog
- Work Management – Jupyter Notebook
- Data Interfaces - OpenDAP
- Analytic Tools: Min, max, sum, diff, average, rms, anomaly, std deviation
- Visualization using time series or spaceplot plot routines

**Climate Impact**
- Project Definition
- Plan for Investigation

**Computational Infrastructure**
- Local Sparc
- SGI Data Migration Facility

**Storage**
- Local systems
- High End Computing (DISCOVER)
- GSFC Science Cloud

POC: laura.carriere@nasa.gov
https://www.nccs.nasa.gov/services/Analytics
Tragedy,

- Large investments by DoD and IC in AI and Machine Learning have, largely, not been leveraged by NASA Earth Science
- Lack of credibility in ML by Earth Science researchers
  - Why does ML generate bogus conclusions?
    - Failure to articulate assumptions & constraints and exceeding them
    - Overtraining
    - Drift over time
    - Failure to recognize when conditions change
  - Earth Scientists have trouble understanding the technology
    - Lack of demonstration of value to motivate them
    - Lack of training in the tools and algorithms
- Partnerships among Machine Learning and Earth Science communities are slow to take root
  - Communications
  - Opportunity
  - Trust
- Data grooming takes huge amounts of time
Myths

• You can do a problem and set it aside for others to use.
  – There is no fire and forget in Earth Science
• Open source is always better than commercial software
  – GIS is a good counter-example
  – Communicating with other Agencies requires arc-GIS capability
• Data must be supplied in same form as collected
  – Stewardship is important, but not useful for analysis
  – Must be able to trace back to authoritative data
• Cannot re-use (trust) anyone else’s work
  – Community acceptance and jurying could create common capabilities
  – Fix the data once and others could use it if they believe it is fixed
Analytic Center Characteristics

• Seamless integration
• Comprehensive catalog
  – Clear applicability (or not)
  – Expert system as an operator aid
• Help in using them
  – YouTube, moodle