Using the NASA Earth Exchange and OpenNEX Platforms

Climate Resiliency Tools & Workflows

Sangram Ganguly
Earth Science Division
NASA Ames Research Center, BAERI

August 4, 2015

Earth Science Big Data Conundrum?

In short

Is Earth Science Data Big?

- Earth Science data is Big Data
- Models and Analytics are core components
- Data center infrastructure is an integral component in delivering data
- High Performance Computing solutions are necessary for actionable intelligence
- Live Data Visualization of large data sets is a pressing need
- More? What are your thoughts?

The Conundrum

- Storing more data
- Accessing large quantities of data faster
- Understanding better what the data tells us (structured vs. unstructured data)
- Integrating efficiently SaaS, Cloud Computing Solutions and Data Access
- Following industry standards

Can We Solve it?

Yes we can

- **Access and sort** data in an efficient manner (deploy open database solutions like Apache Hadoop, Mongo DB, Cassandra, etc.)
- **Improve Legacy Infrastructure** to meet demands of real-time analytics
- **Scale-out, solid-state storage arrays** – fulfill capacity and throughput demands of fast analysis
- **Write** massively parallel applications, distributed jobs, GPU processes
- **Deploy** cloud computing solutions (e.g. Openstack, AWS, Rackspace)
- **Efficient learning algorithms** to extract usable information from overwhelming data
- **Visualize** Data (Highcharts, Tableau, MapBox/Leaflet)
- More? Any thoughts?

NASA EARTH EXCHANGE (NEX).

OVERVIEW

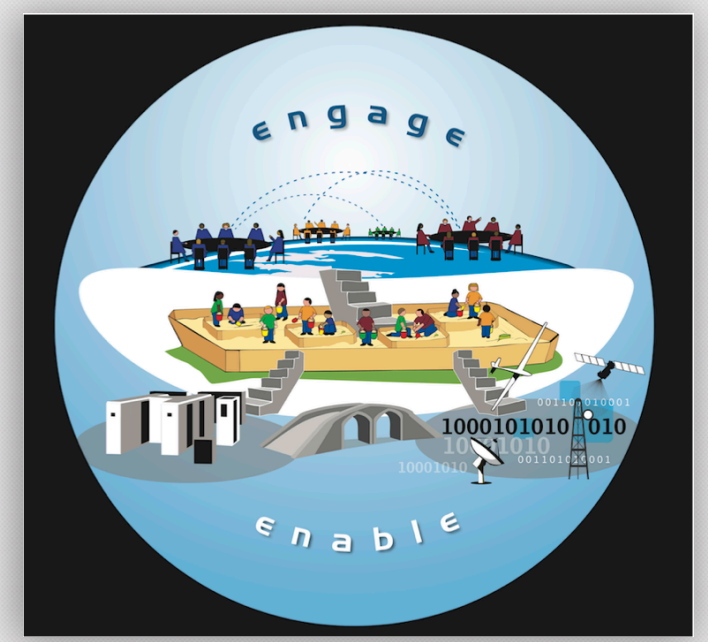
VISION

To provide "science as a service" to the Earth science community addressing global environmental challenges

GOAL

To improve efficiency and expand the scope of NASA Earth science technology, research and applications programs

+ NEX is virtual collaborative that brings scientists and researchers together in a knowledge-based social network and provides the necessary tools, computing power, and data to accelerate research, innovation and provide transparency.



Engage

Network,share & collaborate
Discuss & formulate new ideas
Portal, Virtual Institute

Enable

Rapid Access to data & storage
Access to computing
Access to knowledge/ workflows

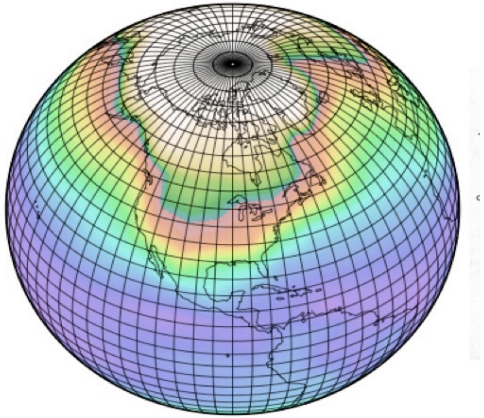
NEX Solutions

- NEX provides access to wide variety of ready-to-use data
- NEX provides the ability to bring “code to data”
- NEX offers capabilities for reproducing science through virtual machines and scientific workflows
- NEX offers state-of-the-art advanced compute capabilities

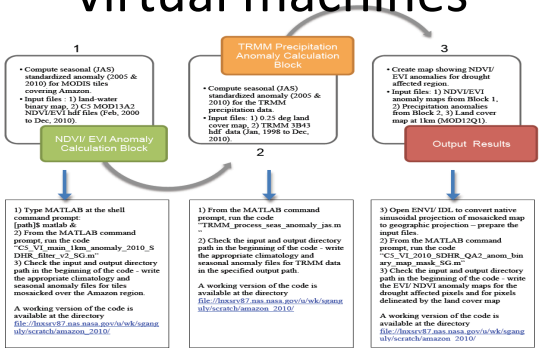
“Science As A Service”

Ready-to-use data

Ready-to-use models

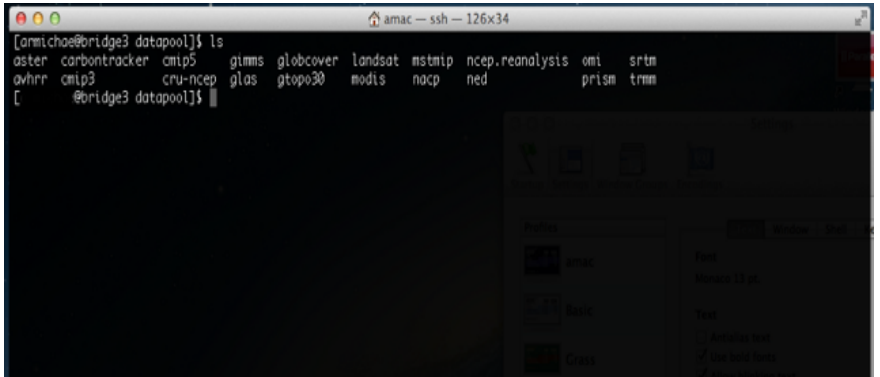


Access to workflows/ virtual machines



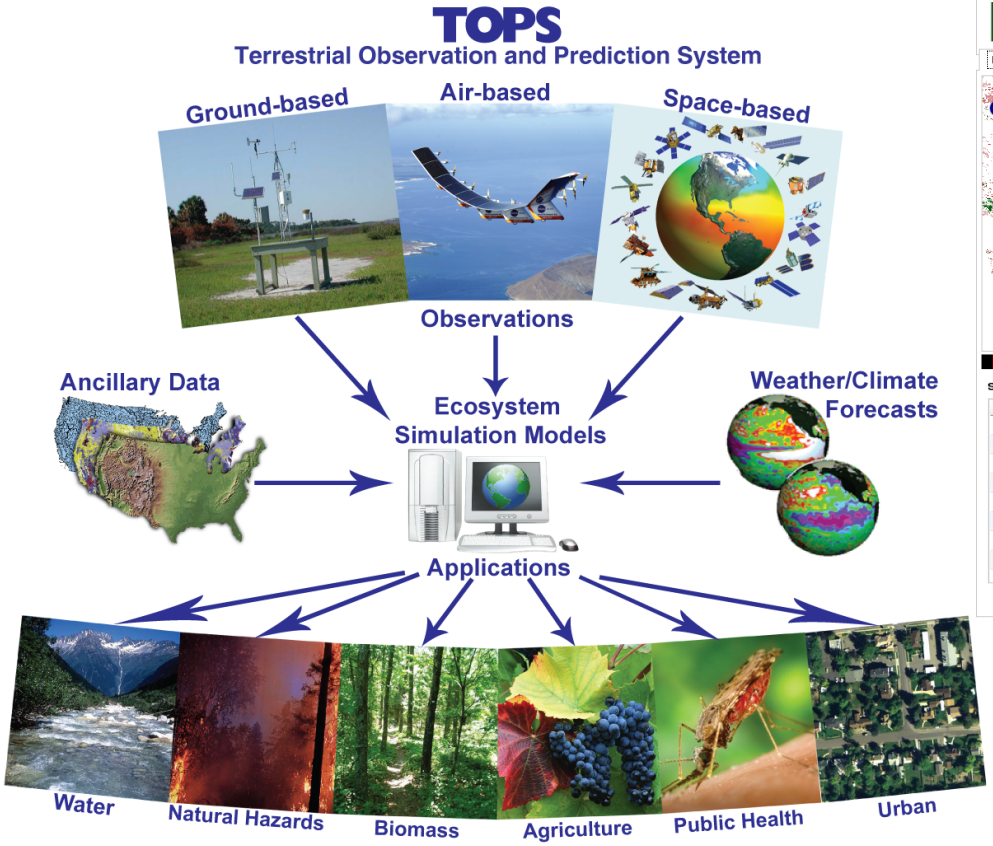
Engage: Web portal

Enable: Terminal

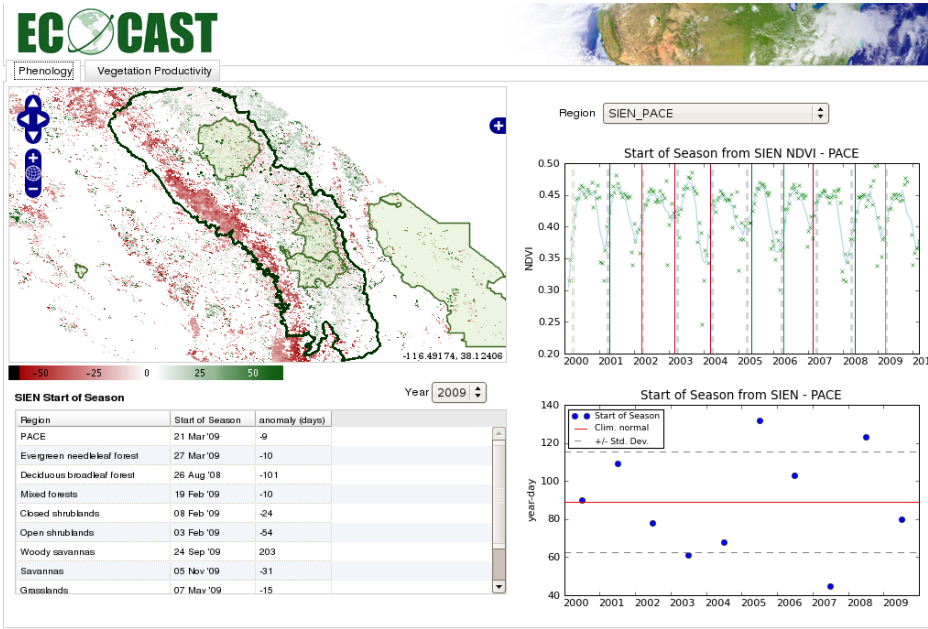


Data to Knowledge

Data-Model Integration



Knowledge Portal

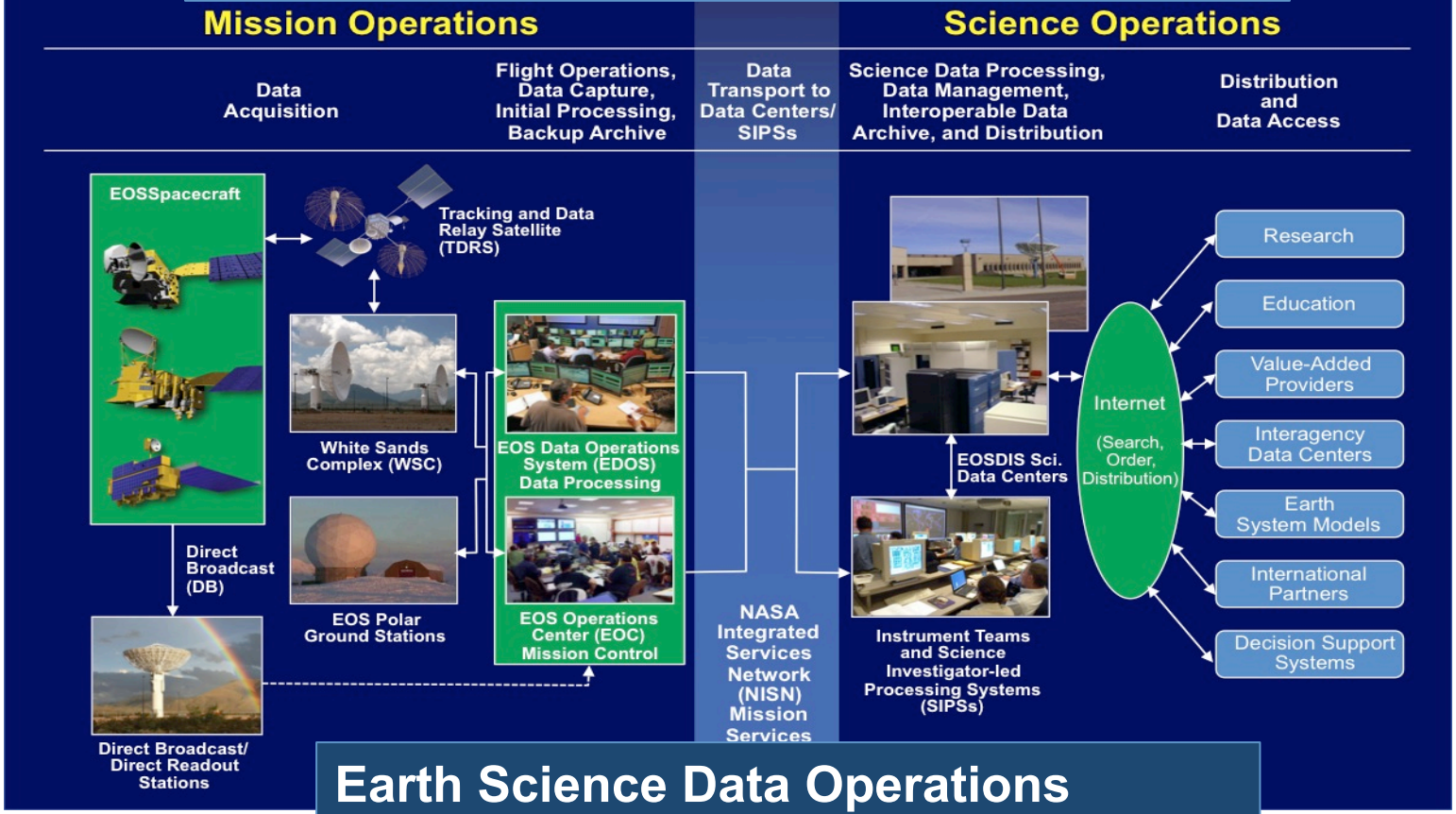


Users often say “show me the data”

Models often produce more data than they ingest

Adapting to new realities

Data volumes, network bandwidth



Earth Science Data Operations

Drivers for an Earth Science Collaborative.

- Researchers spend a major fraction of their time dealing with data (finding, ordering, waiting, downloading, pre-processing...)
- Moving data sets that are getting larger each year over WAN is getting expensive & time-consuming
- Sharing knowledge (codes, intermediate results, workflows) is difficult. Repeated low level IT efforts waste time and resources
- No standard mechanisms for transparency and repeatability
- Culturally – local access is how science is done

NEX Specs…

Portal

- Web Server
- Database Server
- 503 Registered Members

Sandbox

- 96-core server, 264GB memory, will have 320 TB storage
- 48-core server, 128 GB, 163 TB storage

HPC

- 720-core dedicated queue + access to rest of Pleiades
- 181 users/ 44 active (153/40 last year)
- 1.3 PB storage (from 850TB)

Data (>800 TB on & near-line)

Data (450 TB – constantly increasing)

- Landsat (>2M scenes)
- MODIS
- TRMM
- GRACE
- ICESAT
- CMIP5
- NCEP
- MERRA
- NARR
- GLAS
- PRISM
- DAYMET
- NAIP
- Digital Globe
- NEX-DCP30
- WELD

Models/ Tools/ Workflows

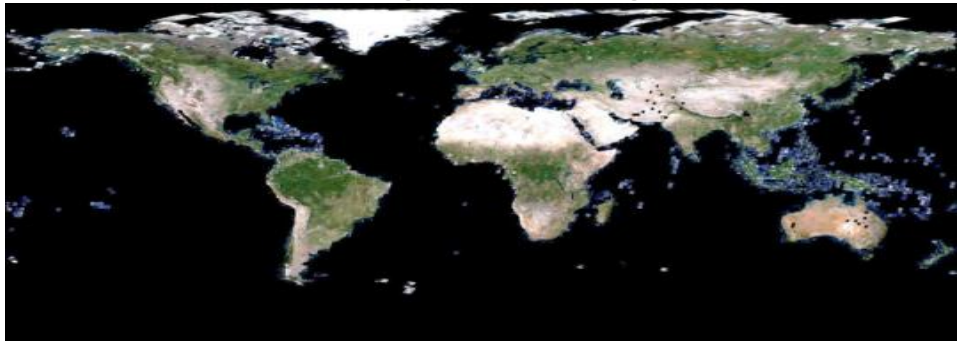
Model Codes

- GEOS-5
- CESM
- WRF
- RegCM
- VIC
- BGC
- CASA
- TOPS
- BEAMS
- Fmask
- LEDAPS
- METRIC

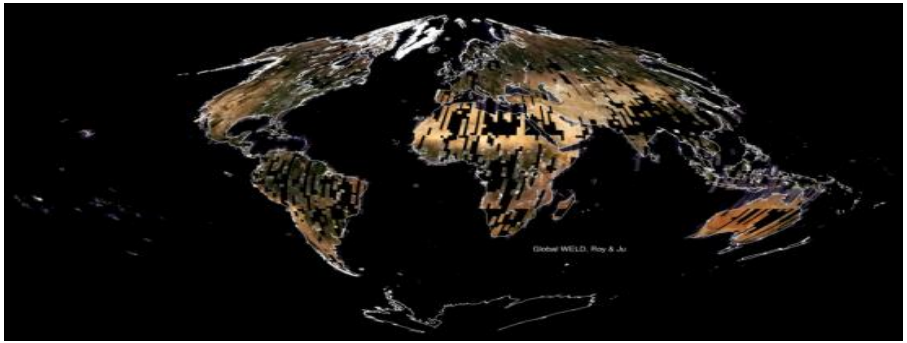
Scale it up

Deployment on NASA's supercomputing resources

From a single scene to global



Mapping global landscapes every month at 30m



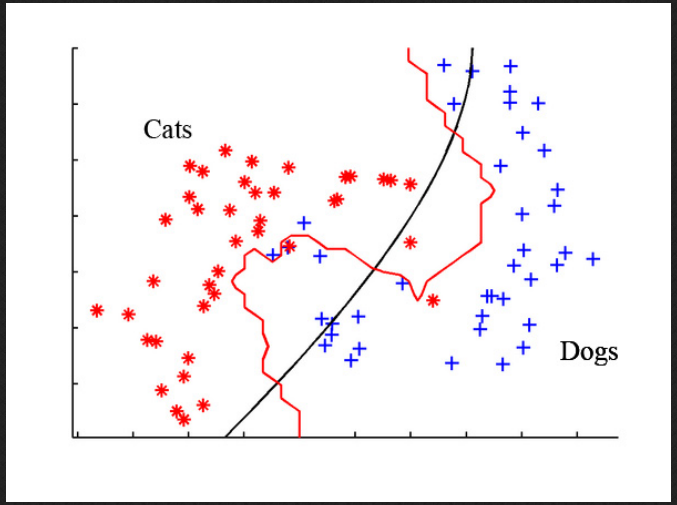
Classes of NEX "Big Data" Projects

- Fully distributed data processing with no inter-process data dependencies
  - Data sizes: 100TB – 5PB
- Data-mining with some inter-process data dependencies
  - Data sizes: 300TB - 2PB
- Analytics and Science Applications
  - Database query systems: 1 – 10TB
- Provenance and knowledge graph queries
  - 100 million to 1 billion triples in 2015
- Climate and ecosystem modeling
  - Computationally intensive time and space data dependent processing
  - 2 – 20TB

Machine Learning & Climate Input Data

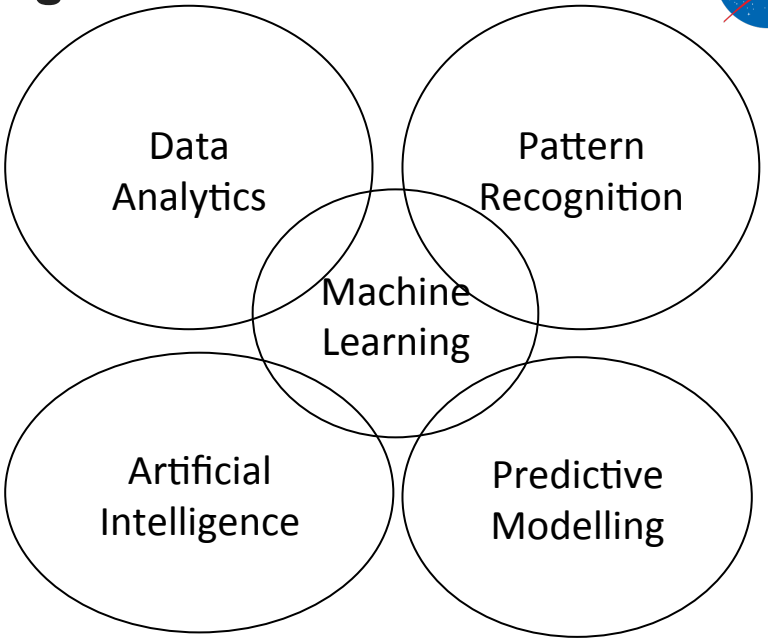
What is Machine Learning ?

- **Machine Learning** is a science that enables us to teach computers to take actions without being explicitly programmed to do so.
- The goal of Machine Learning is Artificial Intelligence.
- Can machines distinguish between cats and dogs?



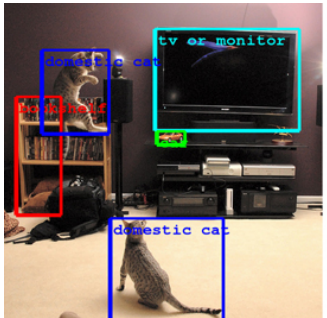
Why do we care about Machine Learning ?

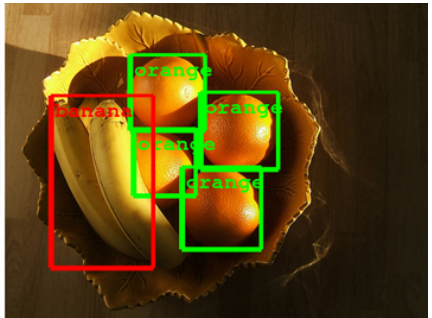
- From Google's search to Facebook's automatic face recognition to Apple's Siri, Machine learning is everywhere.
- Data Analytics and Machine Learning have become synonymous over the years.
- Wherever there is a need for analyzing big data, we need Machine Learning.



What are the questions it can answer?

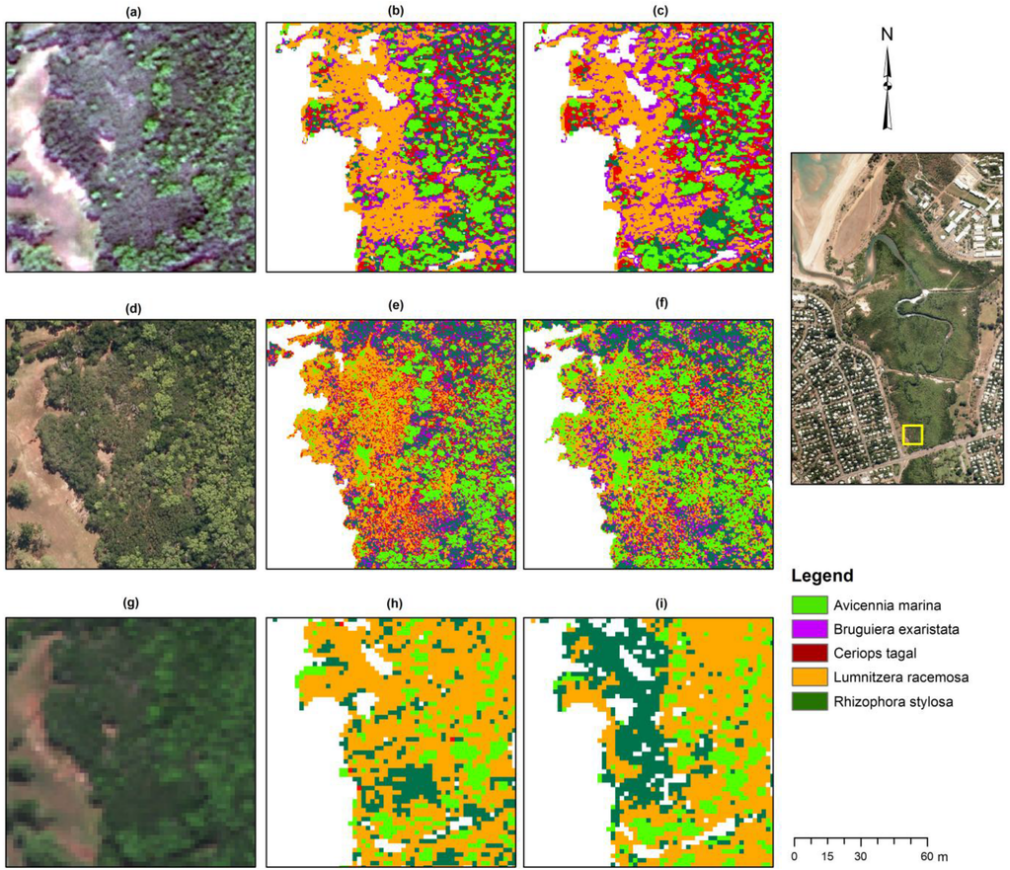
- Does an image contain a particular object?
- Given an image patch, is it possible to label it as belonging to a particular class?

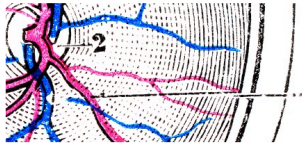




Basic Example?

In a forested or marine landscape, is it possible to segregate different plant species, habitat zones, wind farming zones?





$100,000 • 389 teams

Diabetic Retinopathy Detection

Enter/Merge by

Tue 17 Feb 2015 — Mon 27 Jul 2015 (55 days to go)

Dashboard

Home
- Data
- Make a submission

Information
- Description
- Evaluation
- Rules
- Prizes
- References
- Timeline

Forum

Leaderboard

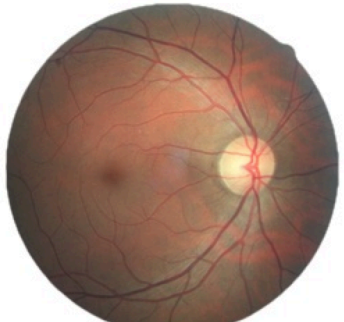
Leaderboard

1. Reformed Gamblers
2. o_O
3. Jeffrey De Fauw
4. Julian de Wit

Competition Details » Get the Data » Make a submission

Identify signs of diabetic retinopathy in eye images

Diabetic retinopathy is the leading cause of blindness in the working-age population of the developed world. It is estimated to affect over 93 million people.



The US Center for Disease Control and Prevention estimates that 29.1 million people in the US have diabetes and the World Health Organization estimates that 347 million people have the disease worldwide. Diabetic Retinopathy (DR) is an eye disease associated with long-standing diabetes. Around 40% to 45% of Americans with diabetes have some stage of the disease. Progression to vision impairment can be slowed or averted if DR is detected in time, however this can be difficult as

Types of Learning?

Supervised Learning

- Learning with a teacher
- Learn a function that maps inputs to outputs using labeled training examples

Unsupervised Learning

- Learn relationships using unlabeled data
- Discover latent patterns in data without supervision

Reinforcement Learning

- Perform a goal in a dynamic and volatile environment
- No teacher supervision
- Driving a vehicle or playing a game against an opponent

Semi-Supervised Learning

- Lies between supervised and unsupervised learning
- Some of the training data are unlabeled
- Goal is to get better predictive performance than either supervised or unsupervised learning alone

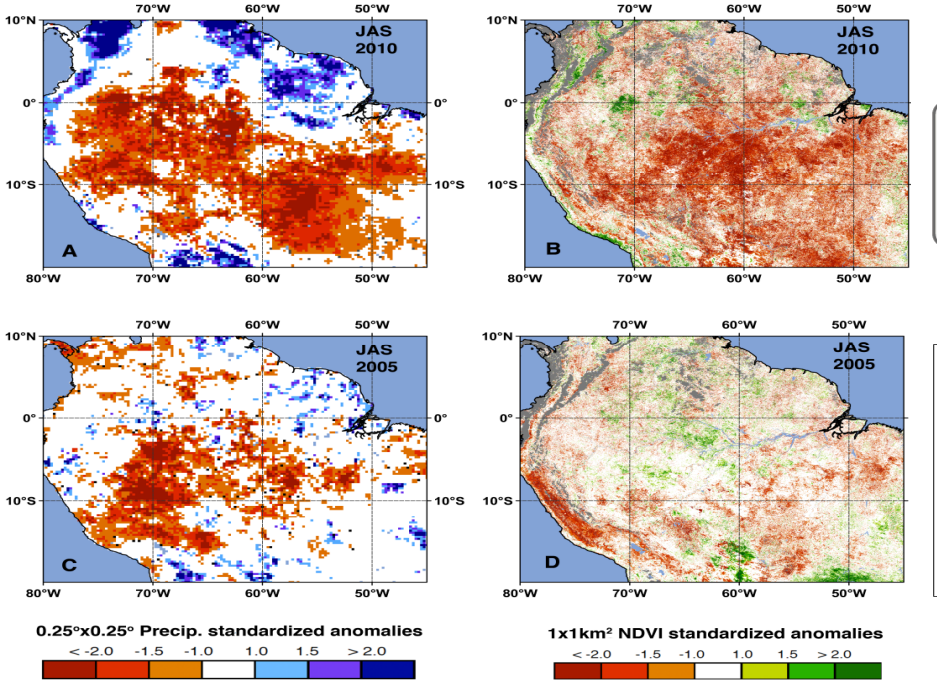
Machine Learning Applications For building Climate Resiliency Tools

Showcasing NEX Projects

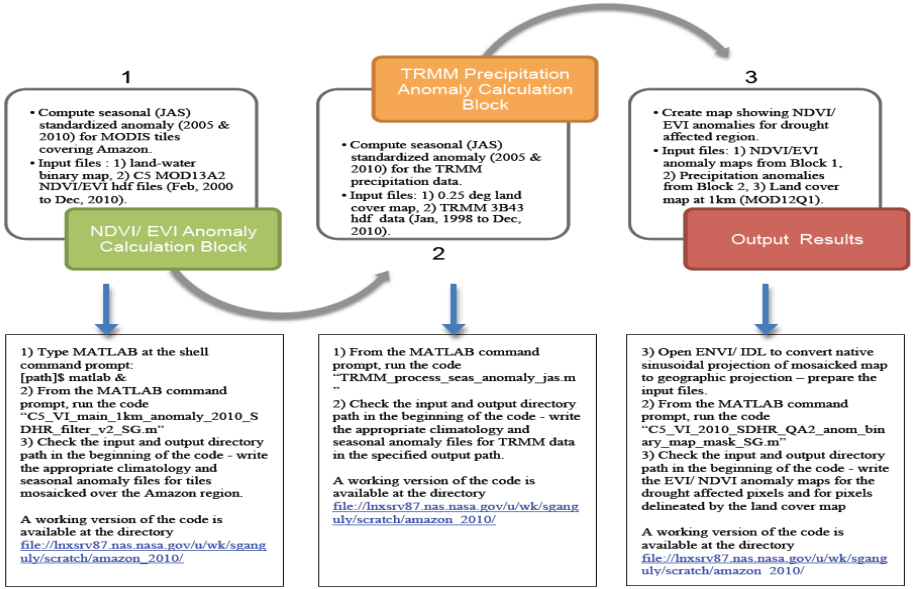
- NEX Satellite Anomaly Workflow
- NEX Global Drought Monitoring
- NEX WELD Processing
- NEX North American Forest Dynamics (NAFD) Processing
- NEX Carbon Monitoring System (CMS) Processing
- NEX Supporting National Climate Assessment (NCA)
- NEX for Agricultural Monitoring

Machine Learning for Anomaly Detection

A tale of two droughts/Amazon 2005 & 2010



TRMM MODIS



Long-term time series data and spatial context – find spatial anomalies in extreme events.

Samanta & Ganguly et al., GRL, 2011
Xu et al., GRL, 2011

Anomaly Detection Workflow.

Global Drought Monitoring, 2012

Total # of Scenes:

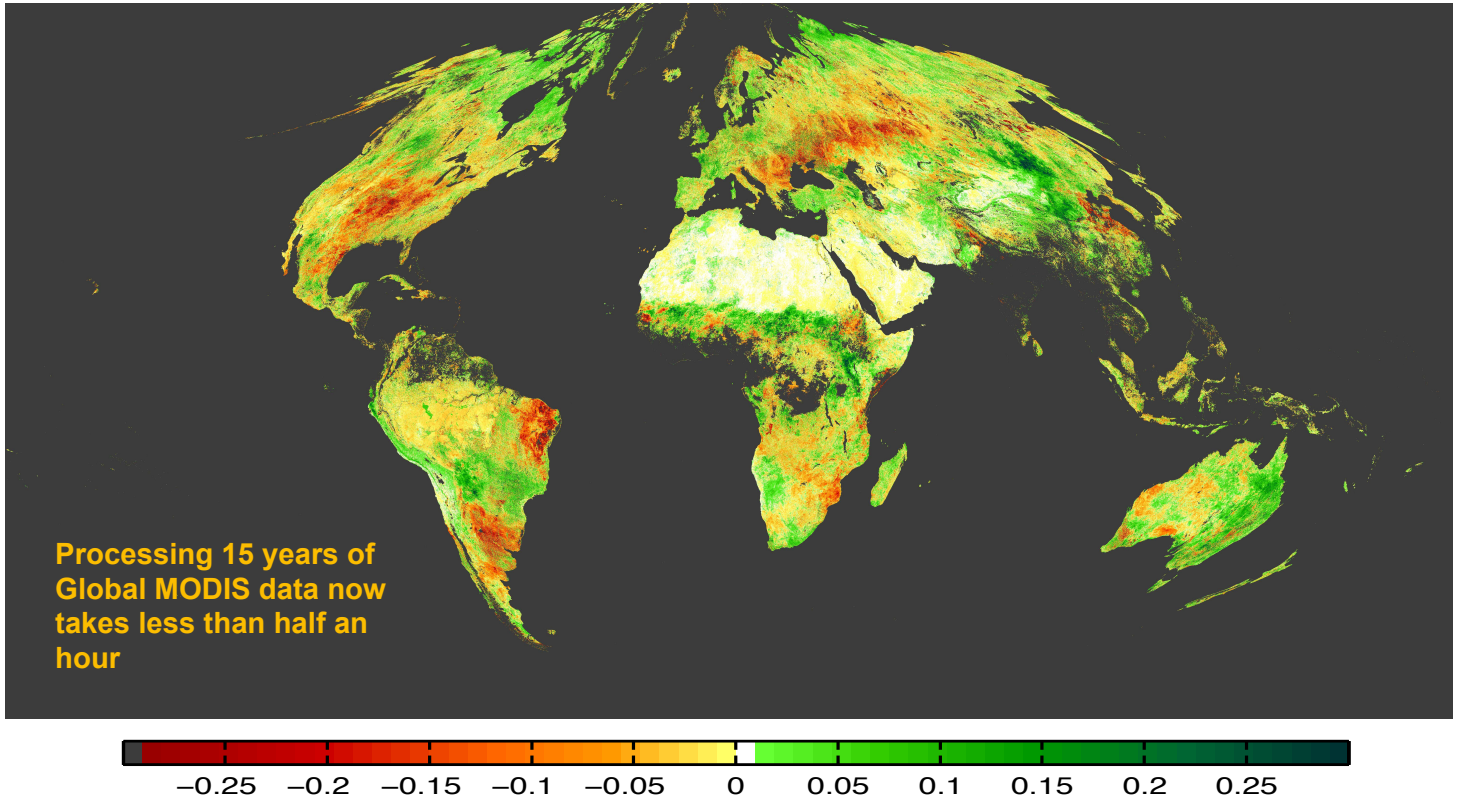
- **1Million for 15 years**

Total Input Data

- **10 TB**

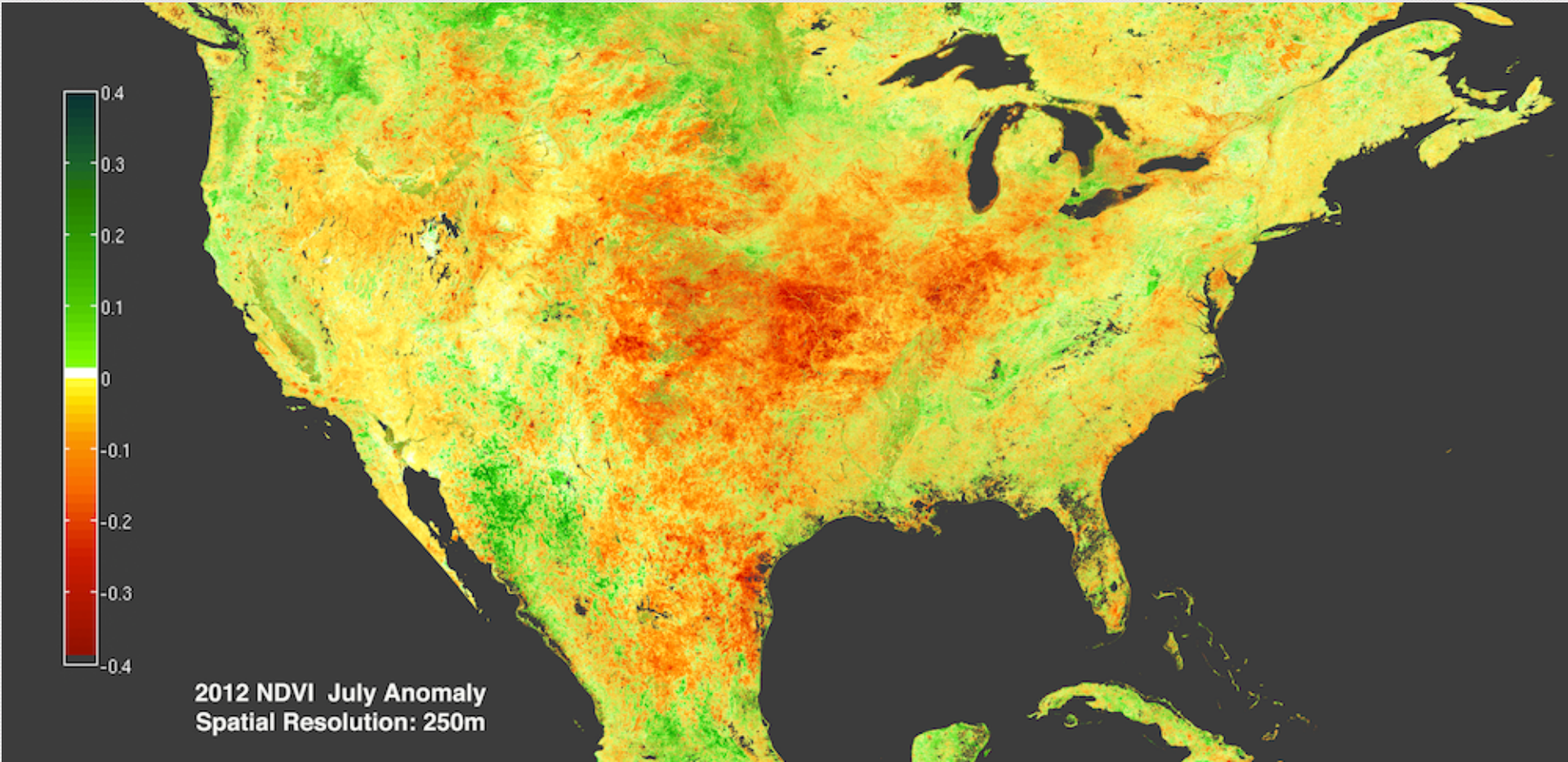
Total Output Data

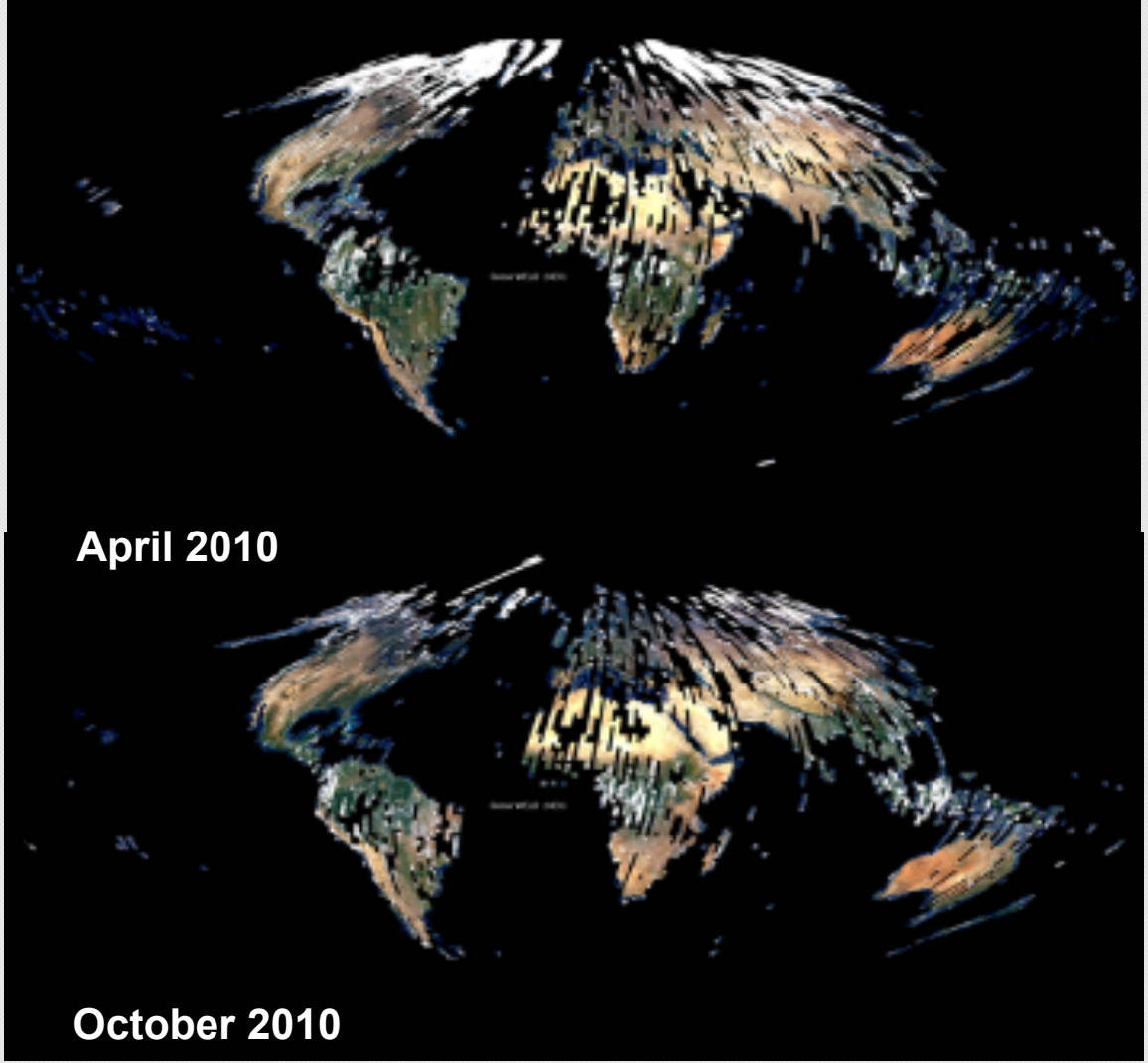
- **50 TB**



Global Drought Monitoring.

2012





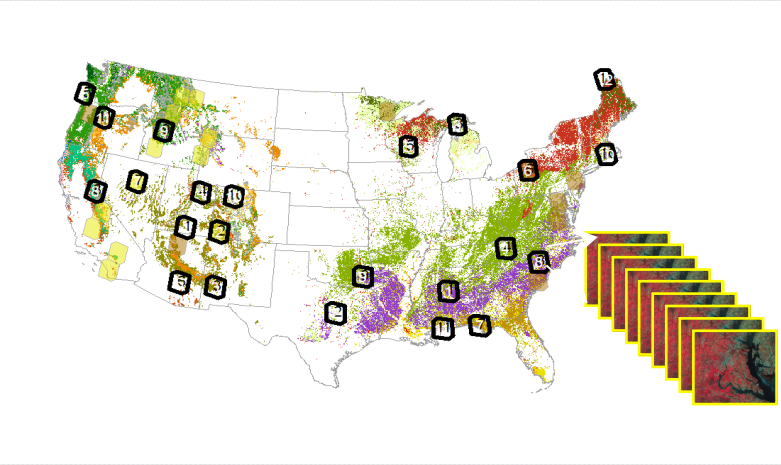
Web Enabled Landsat Data: Going Global, Roy et al.,

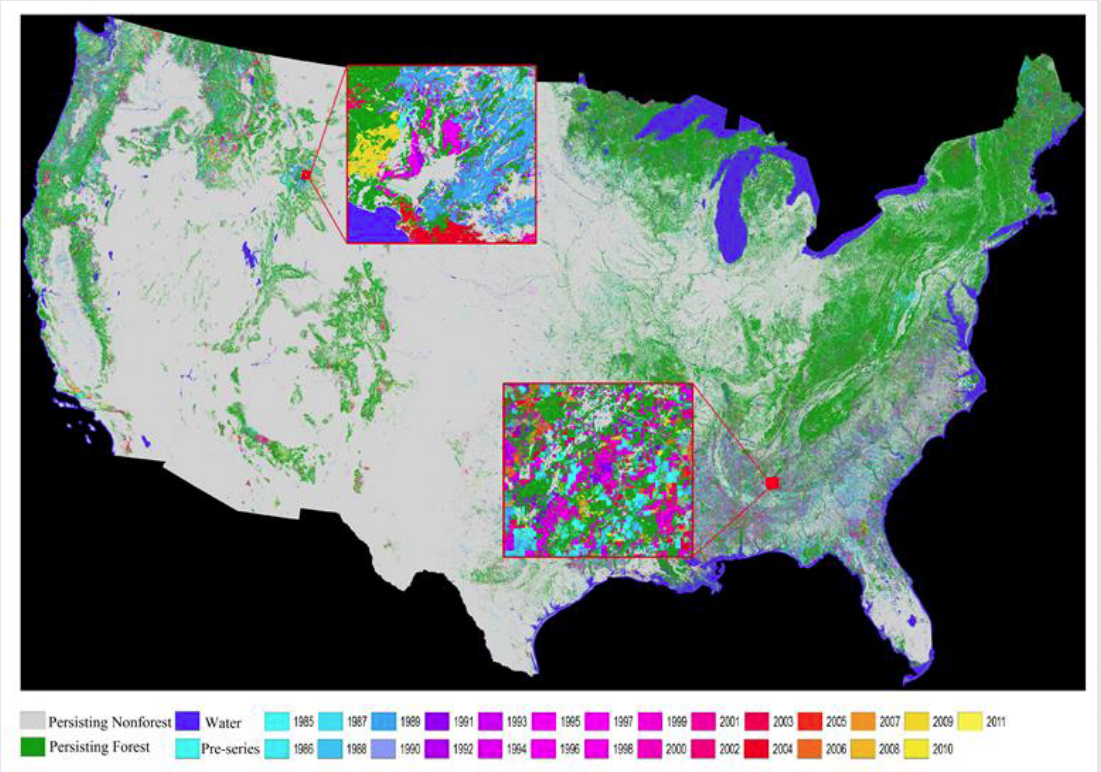
Creating Global Monthly Landsat Composites, 1999 - Present

Takes about 6,000 scenes each month using WELD system

Prototyping land products from Landsat: LAI/FPAR, Albedo

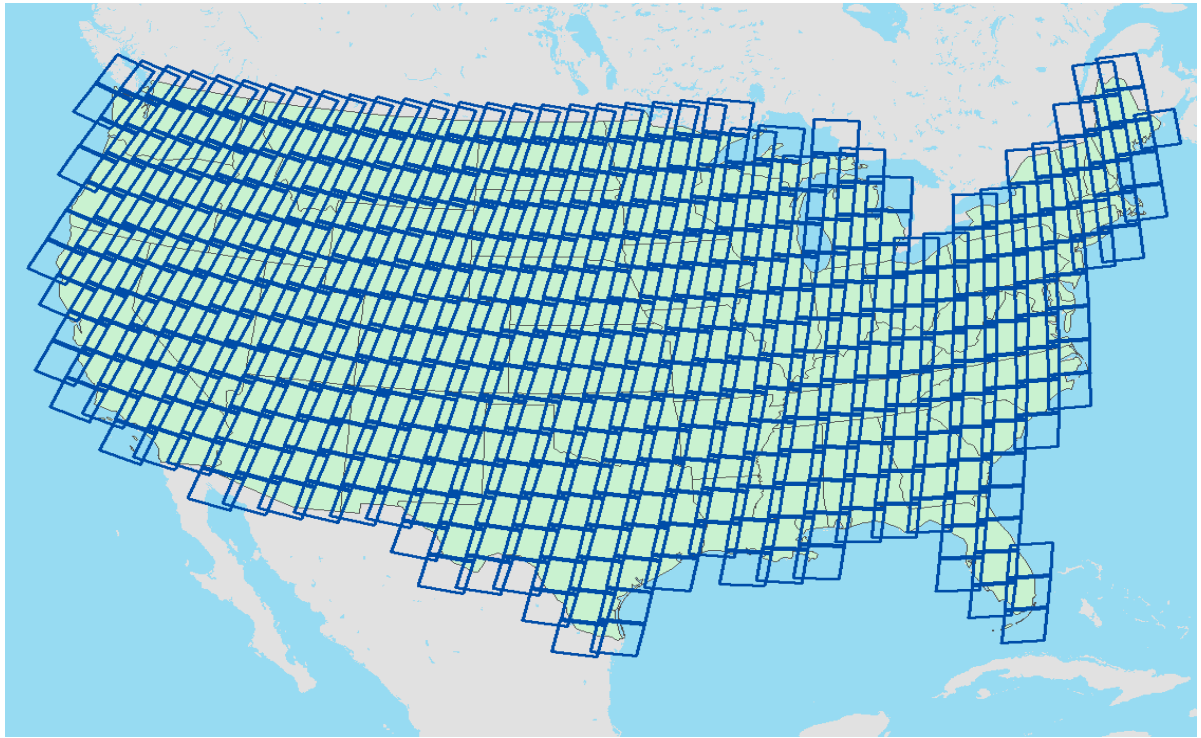
North American Forest Disturbance (NAFD, Goward et al.,)





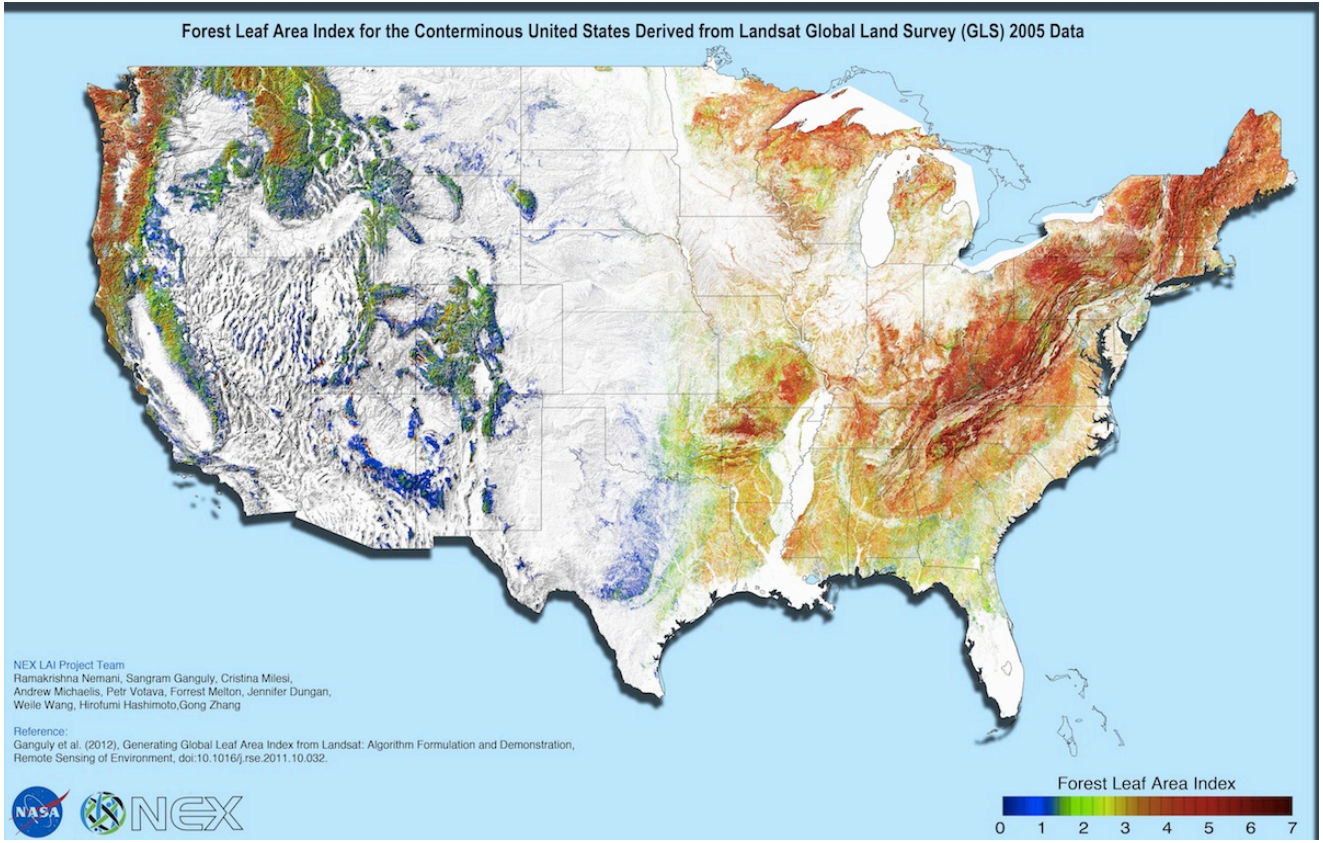
Expanding from 23 samples to Wall-to-wall coverage Processing 96000 scenes from 1985-2010 on NEX

Detecting Forest Disturbance at 30m Spatial Resolution



- 4.3 billon pixels classified as forest
- Polygonize annual disturbance patches and geometric attributes (210,031,105 polygons for US)
- 434 path/rows (scenes) x 29 years x several scenes / year – 16 hours wall time for 12 cores/ node x 1,736 nodes ~ 20,832 cores
- Problem decomposition - scene/node, polygon/core, some polygon overlap across nodes
- Used R packages and custom parallel wrapper

Historical Landsat Analysis.



Landsat Thematic Mapper 1984-2012

Monthly composites of surface reflectances

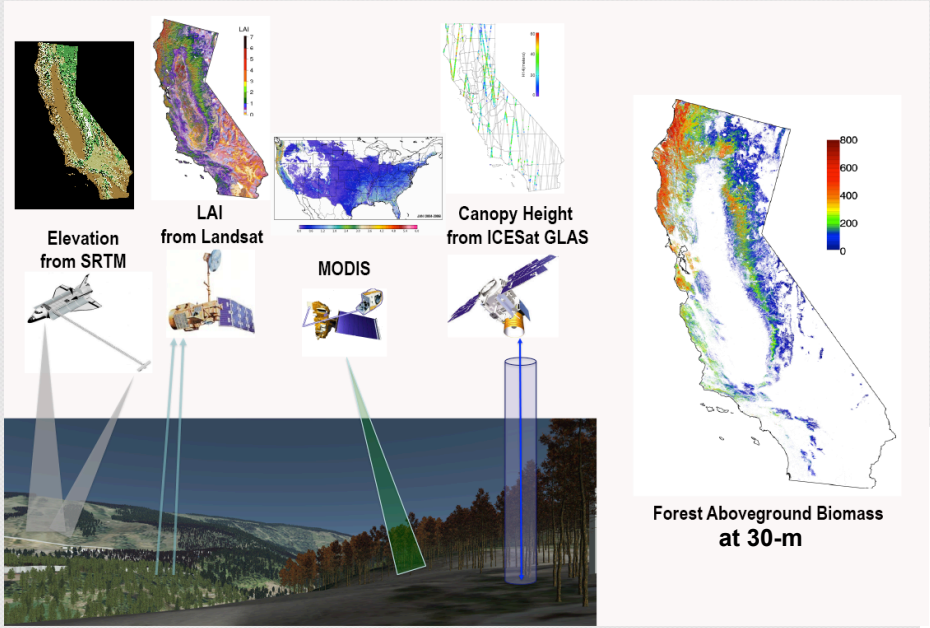
Biophysical products such as LAI

Focus on:

- Land cover changes
- Migration of ecosystems
- High altitude ecosystems
- Forest mortality

Map of Leaf Area Index (LAI) generated using Landsat Thematic Mapper data and a modified MODIS LAI/FPAR algorithm

NEX supporting the **Carbon Monitoring System (CMS)**.

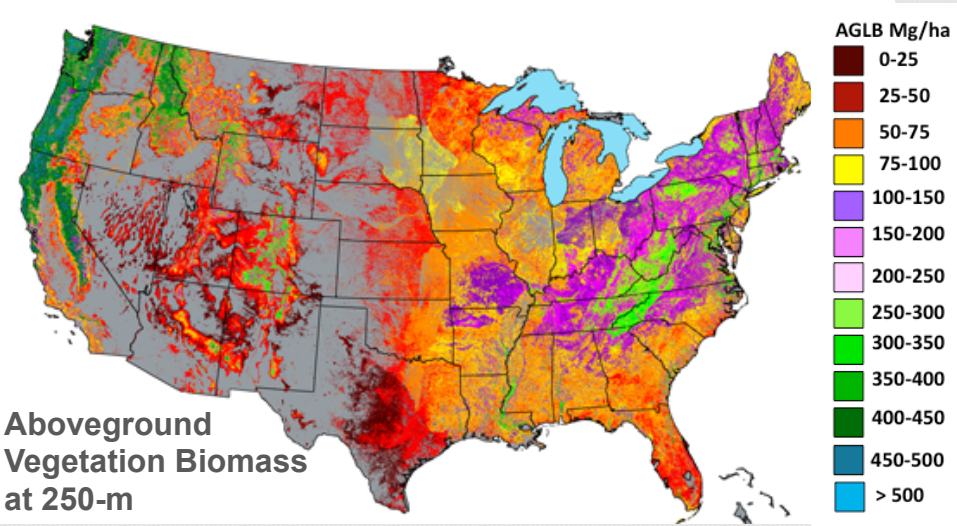


Forest Aboveground Biomass
at 30-m

Several Process-based models and machine learning algorithms are used to estimate total carbon sequestration potential for the U.S. forests.

Creating 3D vegetation from multiple sensors

Saatchi & Ganguly et al.

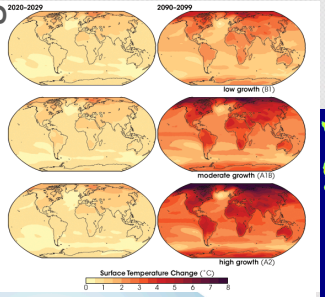


Aboveground Vegetation Biomass at 250-m

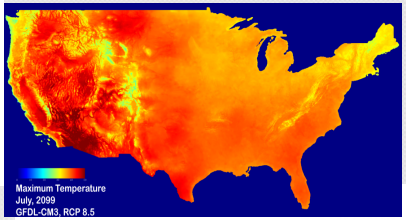
NEX supporting the National Climate Assessment (NCA).

Promoting consistency, repeatability, and transparency in global change science

Climate modeling

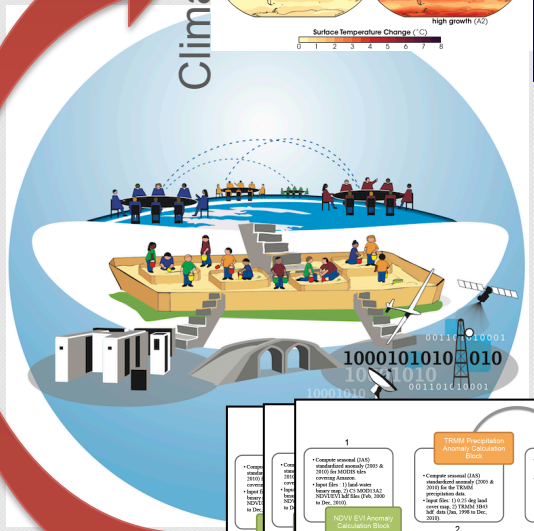


Downscaling

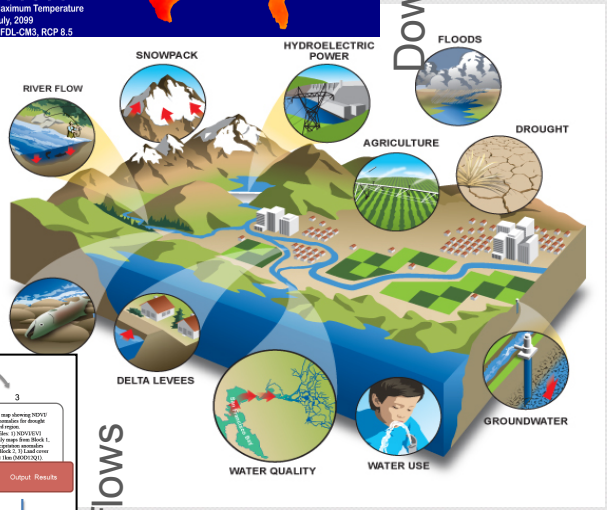


Machine Learning Model

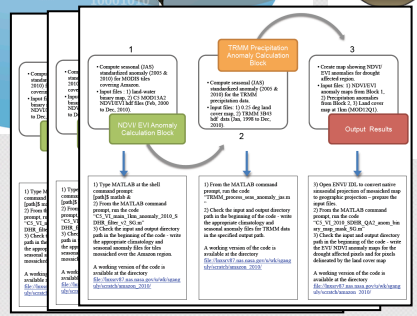
Every 4 years



Modeling impacts

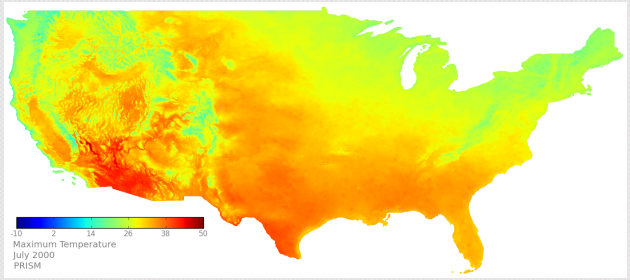


Workflows

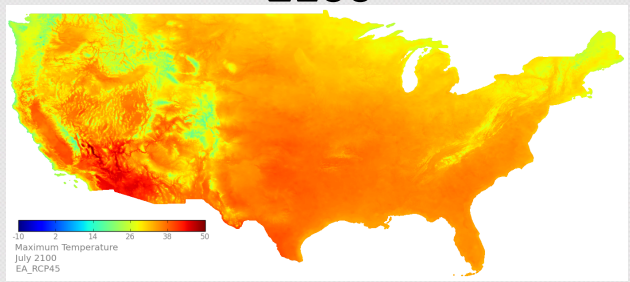


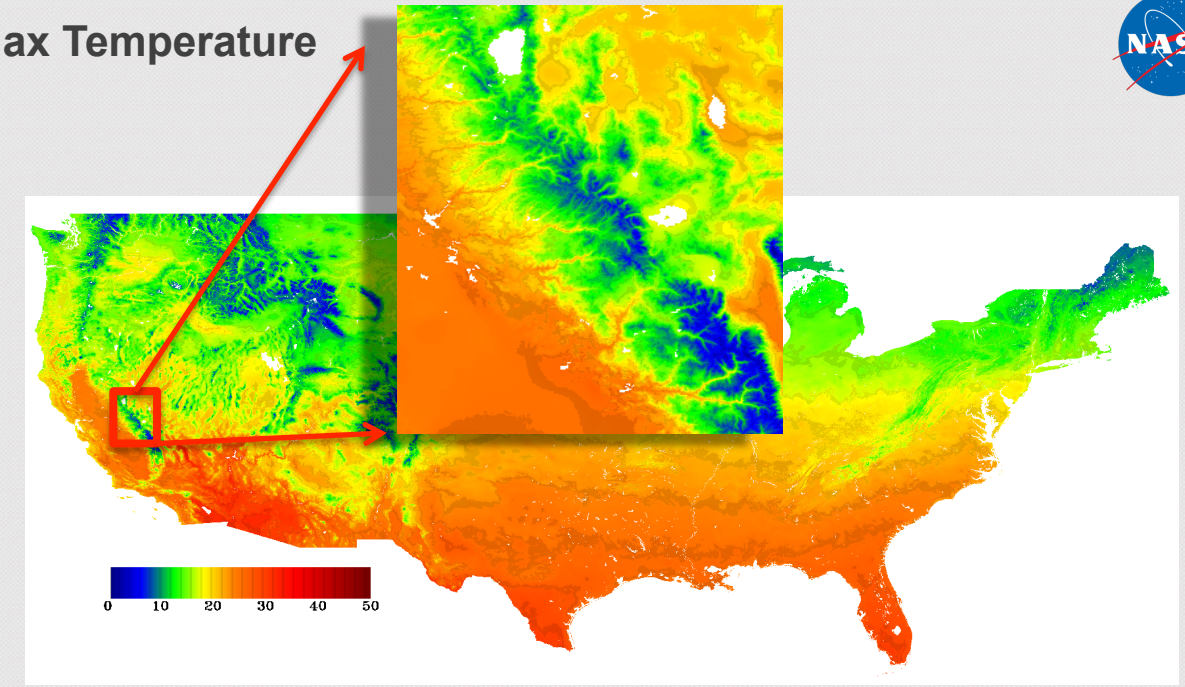
July Max Temperature

2000



2100





Input: 35 CMIP5 models
Downscaling method: BCSD
Temporal Resolution: Monthly
Spatial Resolution: 800m
Temporal Resolution: 2006-2100
PRISM data for bias correction

Output variables:
Ave Max Temperature
Ave Min temperature
Total Precipitaiton
Ave Humidity
Ave Solar Radiation

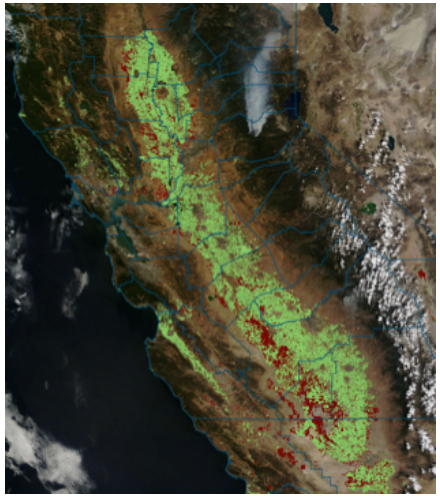
Individual model outputs
Ensemble means
Percentiles
Volume: 22TB

Distribution of the downscaled data

From: GSFC/NCCS
Format: Earth System Grid (ESG) API from Google Earth Engine

Mapping of Crop Water Requirements and Drought Impacts on Ag Production

- Prototyping of workflows for CA Department of Water Resources for mapping crop water requirements and drought impacts on ag production
- Extraction and analysis of VI timeseries for 200,000+ agricultural fields in CA Central Valley
- Distributed processing of 18,400 Landsat scenes used in prototyping for California / calculation of 10 year baseline
- NEX workflow supports scaling for other regions / states
- Data featured in 5 pg poster in Nat Geo story on drought in the West (Oct 2014)



Sept 22, 2014
Central Valley
Summer Conditions
(June 1 – Sept 22)

OpenNEX

NASA EARTH EXCHANGE

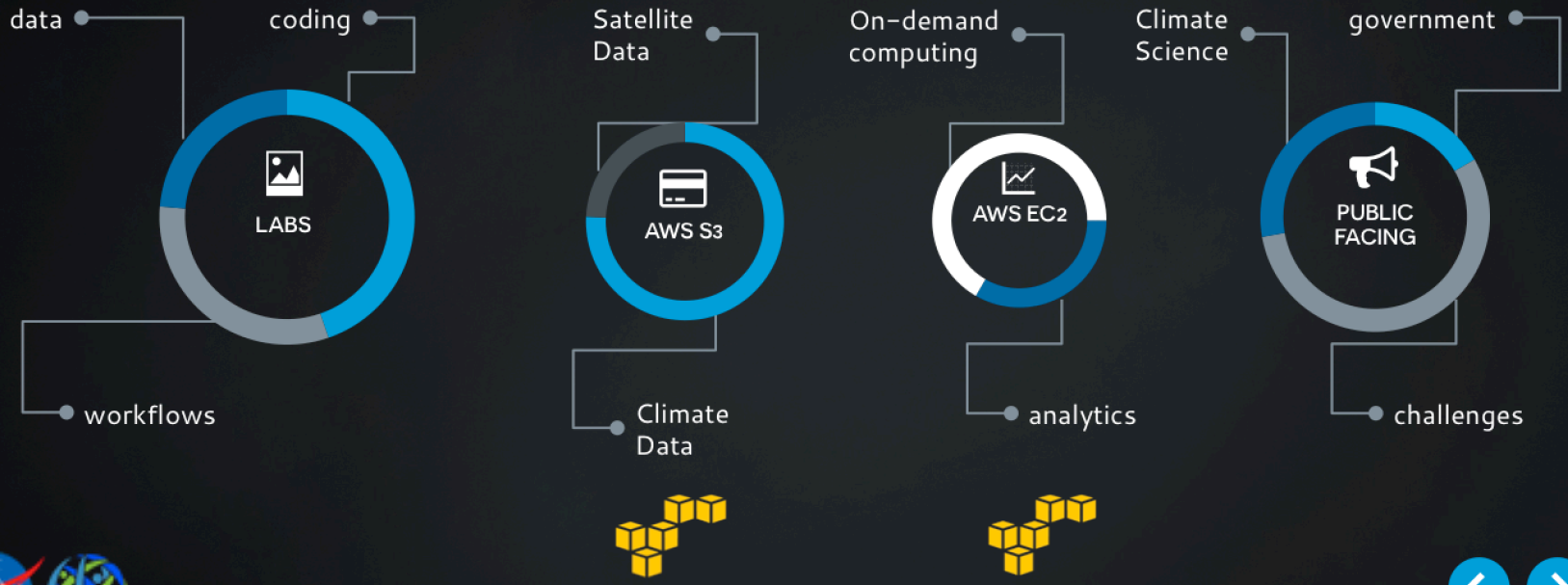
Development Team

- Ramakrishna Nemani
- Andrew Michaelis
- Sangram Ganguly
- Petr Votava

Previous Year Effort

CORE FEATURES

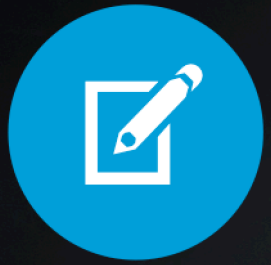
Not just a portal – much more !



Previous Year Effort

OpenNEX Workshop 2014

Why this workshop?



DATA

- Access satellite data like MODIS, Landsat in your VM
- Access historical and projected climate data
- Populate data based on public demand

VIRTUAL LABS

- Build your own VM or use pre-built ones
- Access on-demand computing
- Demos for hands-on labs

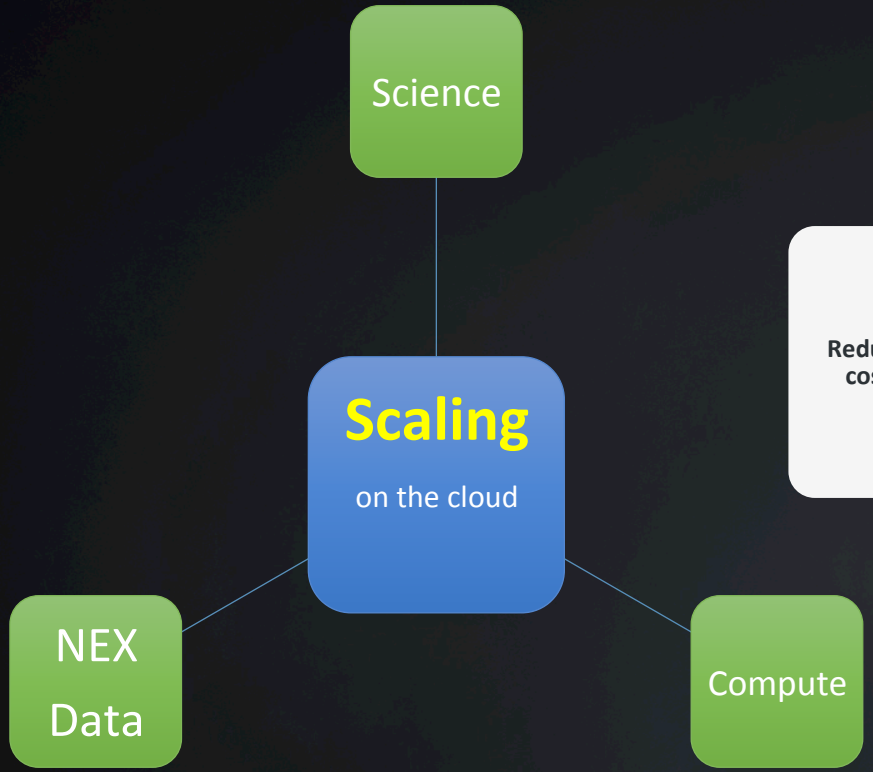
COMPETITION

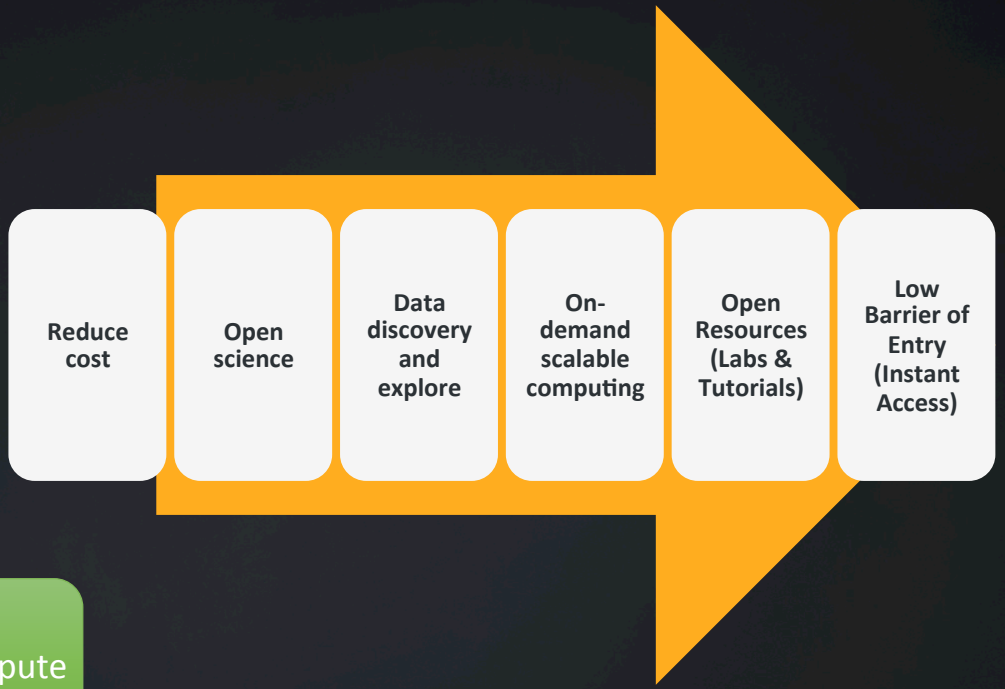
- Design concepts and solutions to enable climate resilience
- Build web or mobile applications
- Win prizes

LECTURES

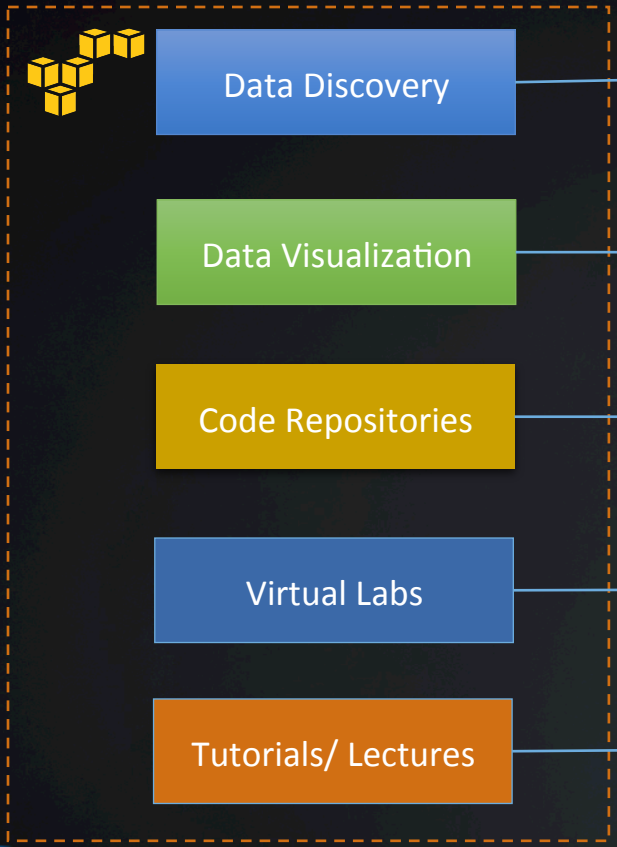
- Live lectures from climate and remote sensing experts
- Access to pre-recorded lectures
- Invite and Share

WHY IS SCALING IMPORTANT ?





OpenNEX New Features



OpenNEX Stack

AWS

FRONT END

HTML
CSS
JS
Bootstrap
AJAX
jQuery
HIGHCHARTS
Leaflet

BACK END

Microsoft SQL Server
Microsoft .NET
C#
IIS
redis

AWS STACK

Compute
- EC2
- Lambda

Networking
- VCP
- Route 53

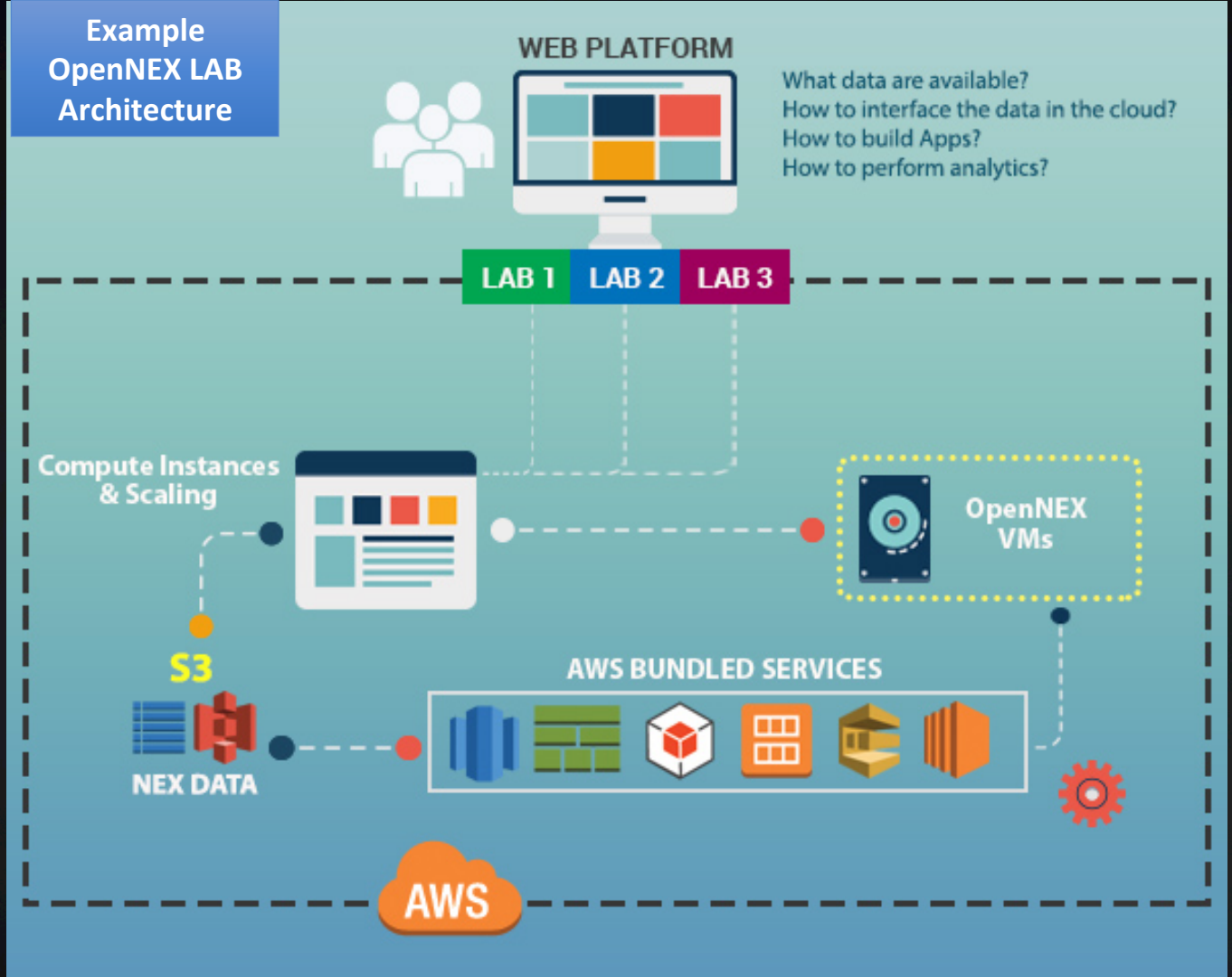
Storage
- S3
- EFS
- CloudFront

App Services
- SQS
- SES
- API Gateway

Database
- DynamoDB
- ElastiCache

Deploy & Mgmt
- CloudFormation
- CodeCommit

Example OpenNEX LAB Architecture

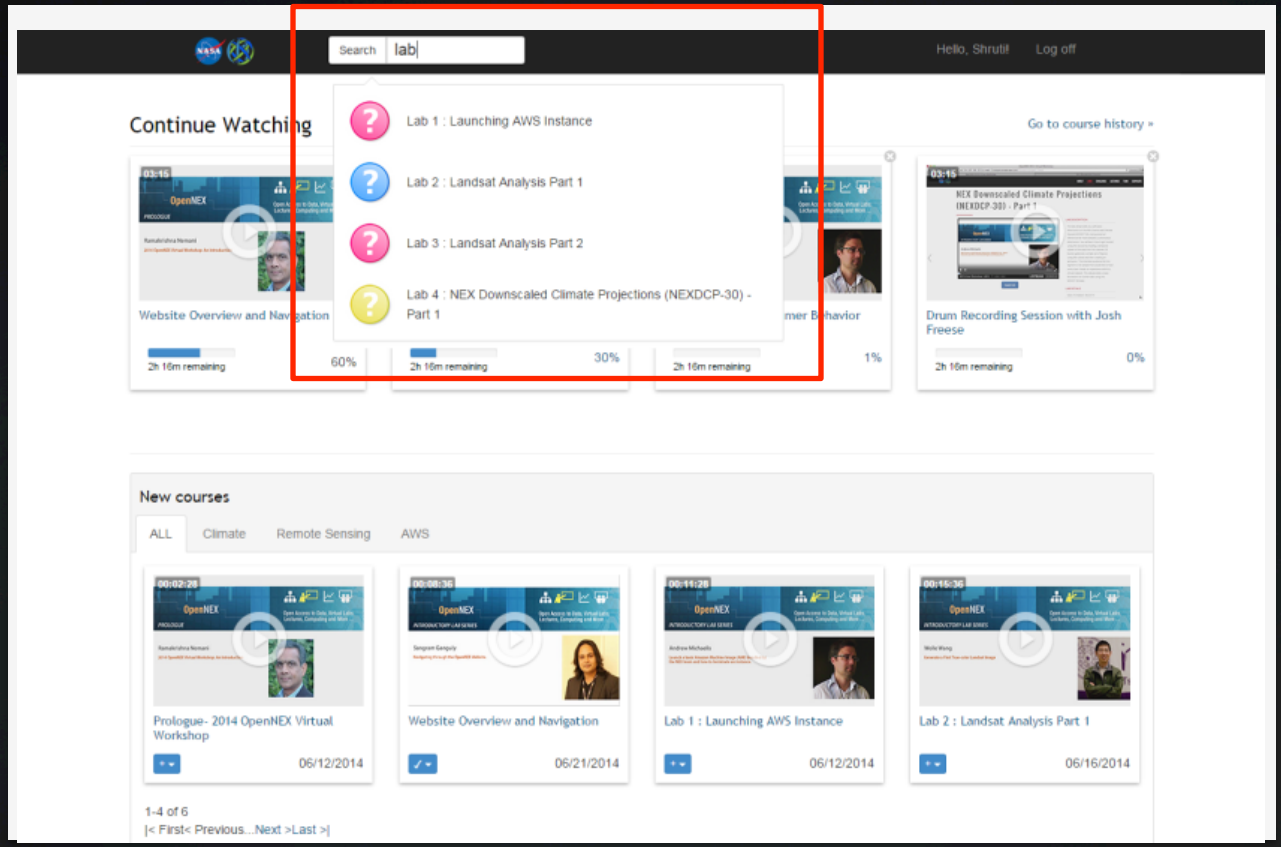


OpenNEX User Features

- Wide variety of Lectures/Courses **categorized by topics**
- **Playlists** – User can create their own Playlists based on their taste
- **History** – User can access all the lectures they have watched earlier
- **Progress Tracking** – User can resume partially watched lectures from where they left

Useful for knowledge capture, usage and analytics

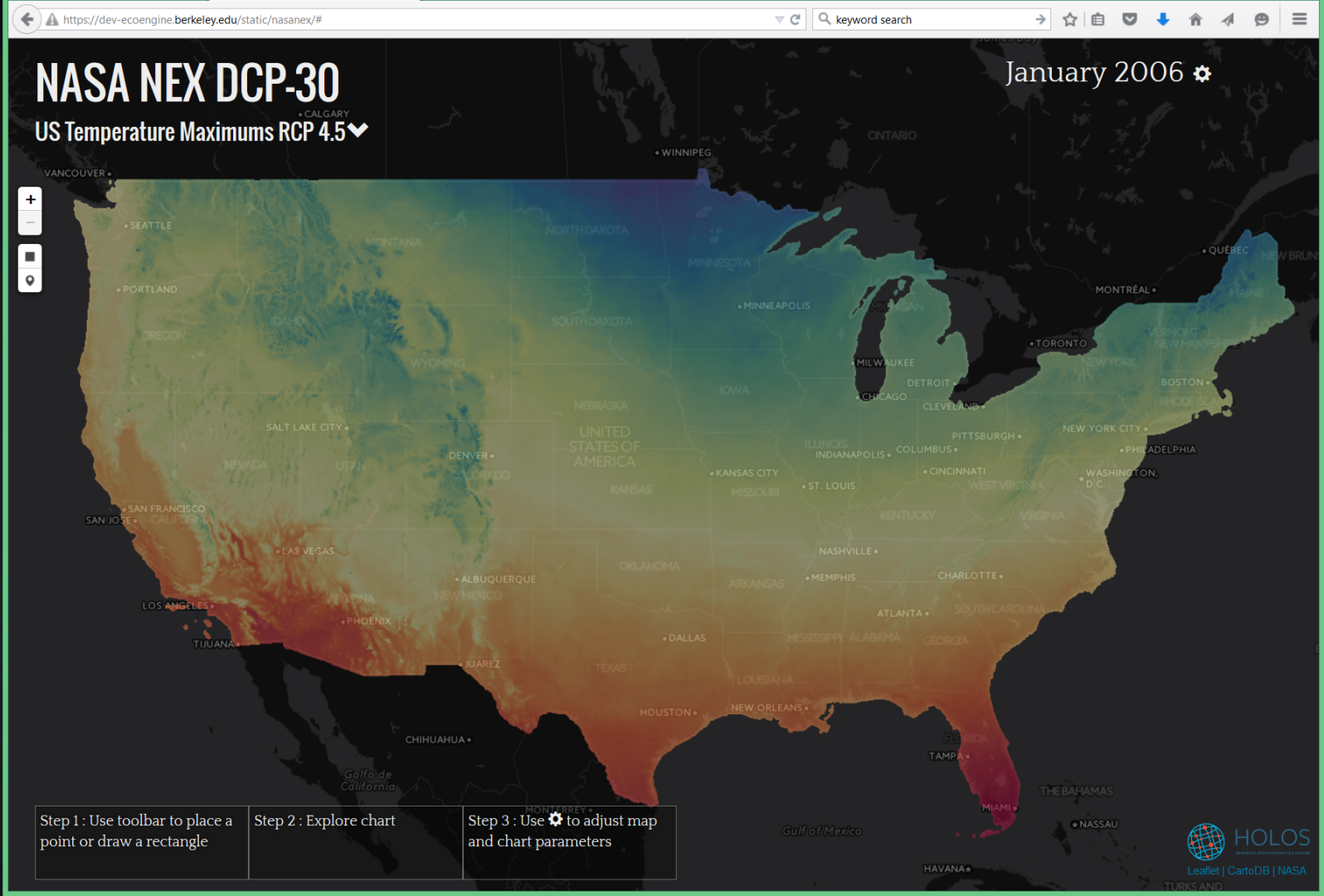
Search / Auto-suggestions



Data Visualization Interface

Features (e.g.) :

- Leaflet
- TileStache
- D3
- NVD3
- Highcharts
- Data Engine
- Rest APIs
- Event Database



Interactive Charting & Maps

https://dev-ecoengine.berkeley.edu/static/nasanex/#

keyword search

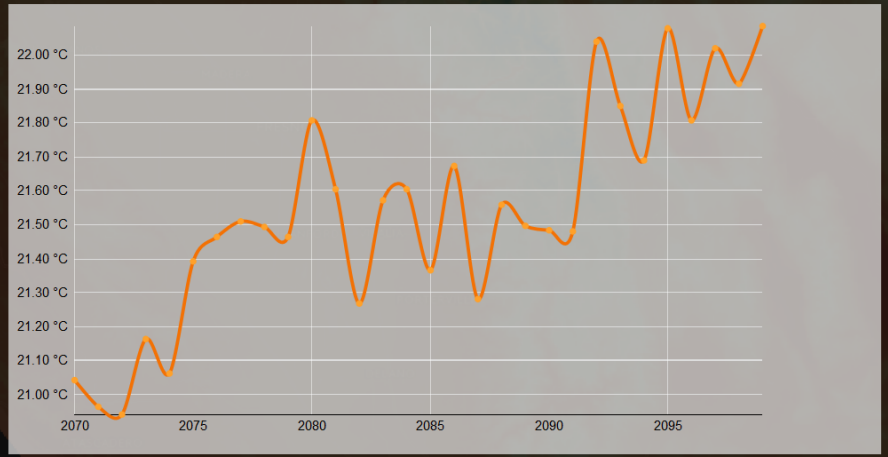
NASA NEX DCP-30

US Temperature Maximums RCP 6.0

January 2070

2006 — 2099

Time series



Step 1 : Use toolbar to place a point or draw a rectangle

Step 2 : Explore chart

Step 3 : Use to adjust map and chart parameters

Model: Ensemble Average Scenario: RCP 6.0

Search & explore a live index of geospatial data

Save your searches

Arrange finds into collections

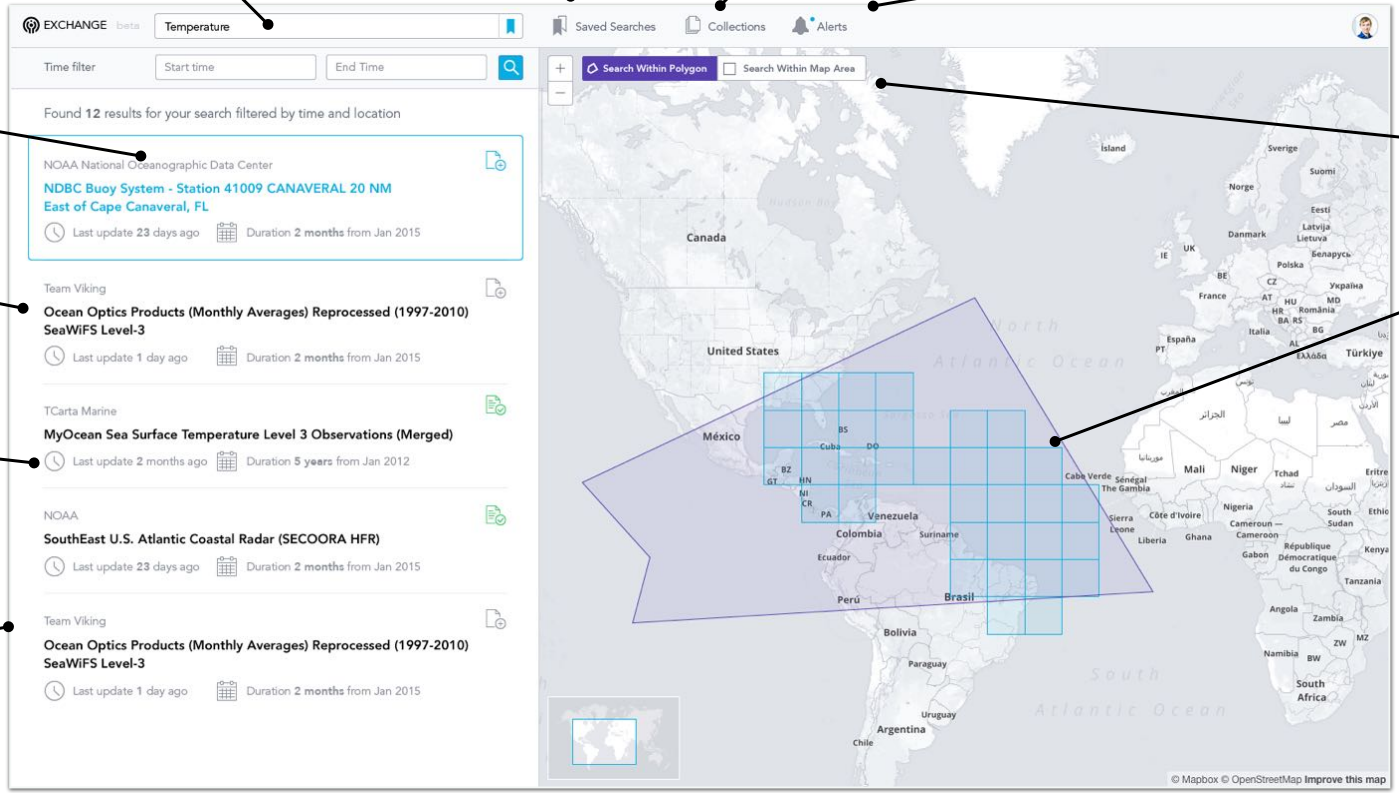
Set an alerts to get relevant content notifications

Dataset detail pages with metadata

Search ranking

Temporal coverage, resolution, frequency

See public and private data in one system



Polygon and map filters

Adaptive spatial coverage

Expandable framework for new features

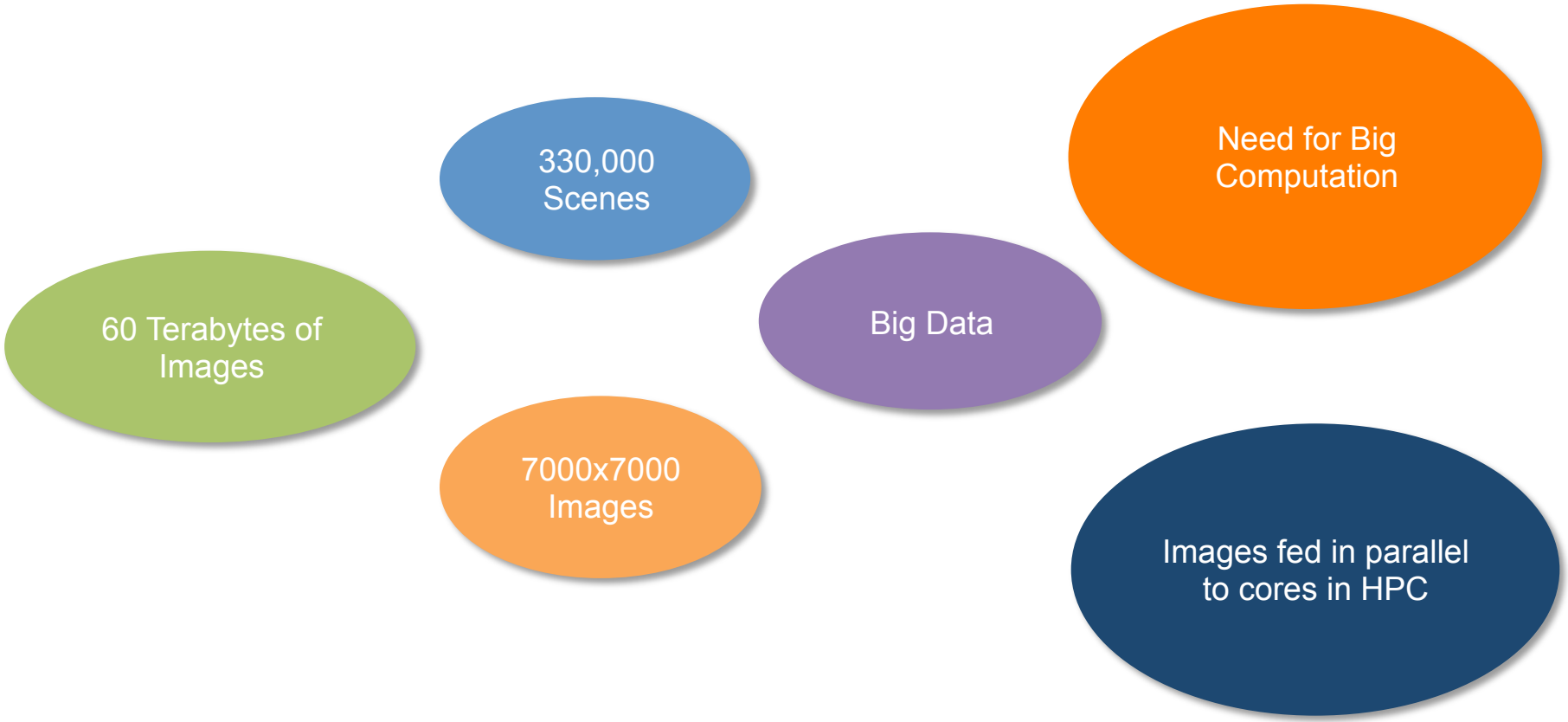
Very High Resolution Satellite Image Classification

Showcasing NEX Projects

- NASA Carbon Monitoring System (CMS) NAIP Data Application

- NASA Advanced Information Systems Technology (AIST) Program Application

NAIP – Deriving Tree-cover from 1-m Imagery for CONUS.



Current End-to-end Processing Time (California with 11,000 scenes) -> 48 hours

Problem and Motivation

Quality of data affected by data acquisition, pre-processing and filtering.

Significant inter-class overlaps and often hard to distinguish between classes.

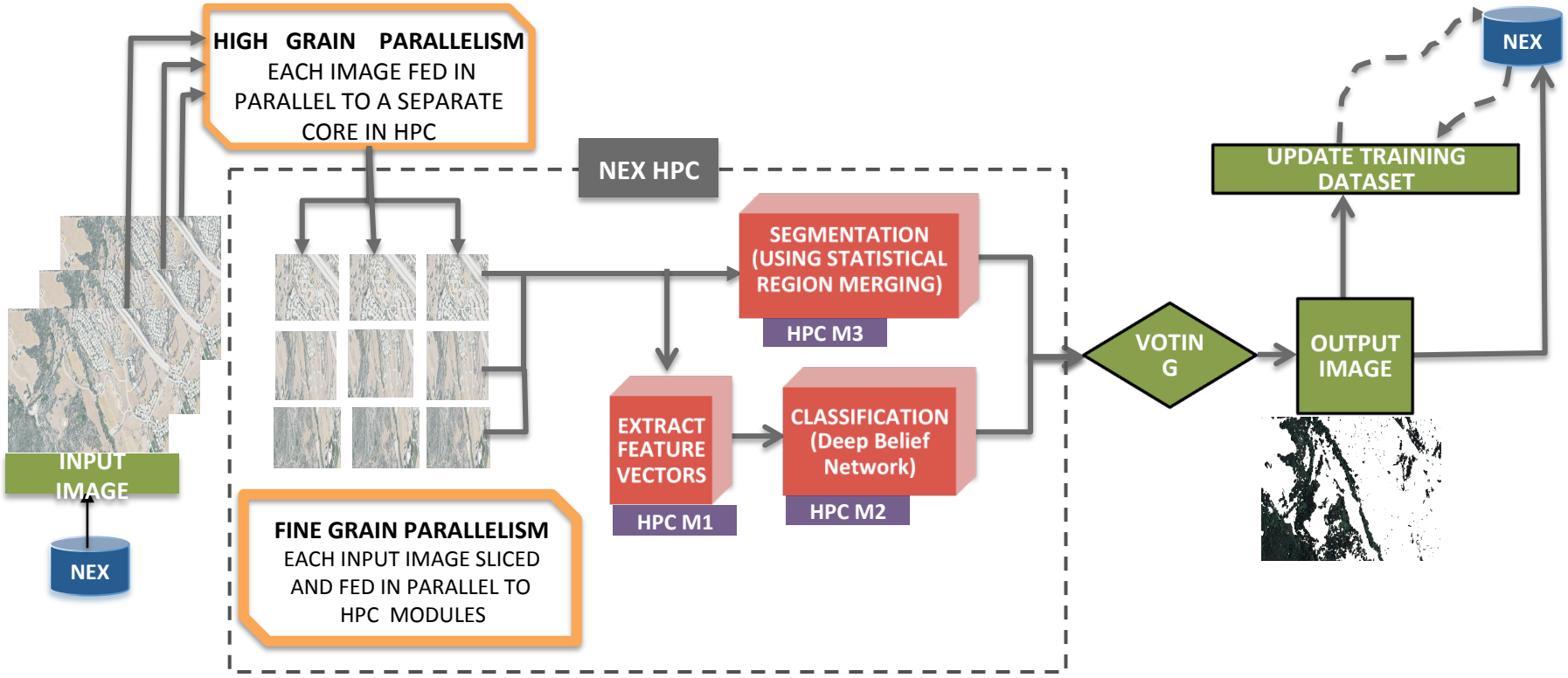
Tree cover delineation is a *hard* problem

Need to harness strong discriminative features and efficient learning algorithm.

Accuracy of present algorithms is low and there is a pressing need to create high resolution land cover maps.

We create a learning framework by combining *unsupervised segmentation* and *deep learning based classification* which produces state-of-the-art results.

NAIP Processing Architecture



HPC M1 **HPC Module 1**

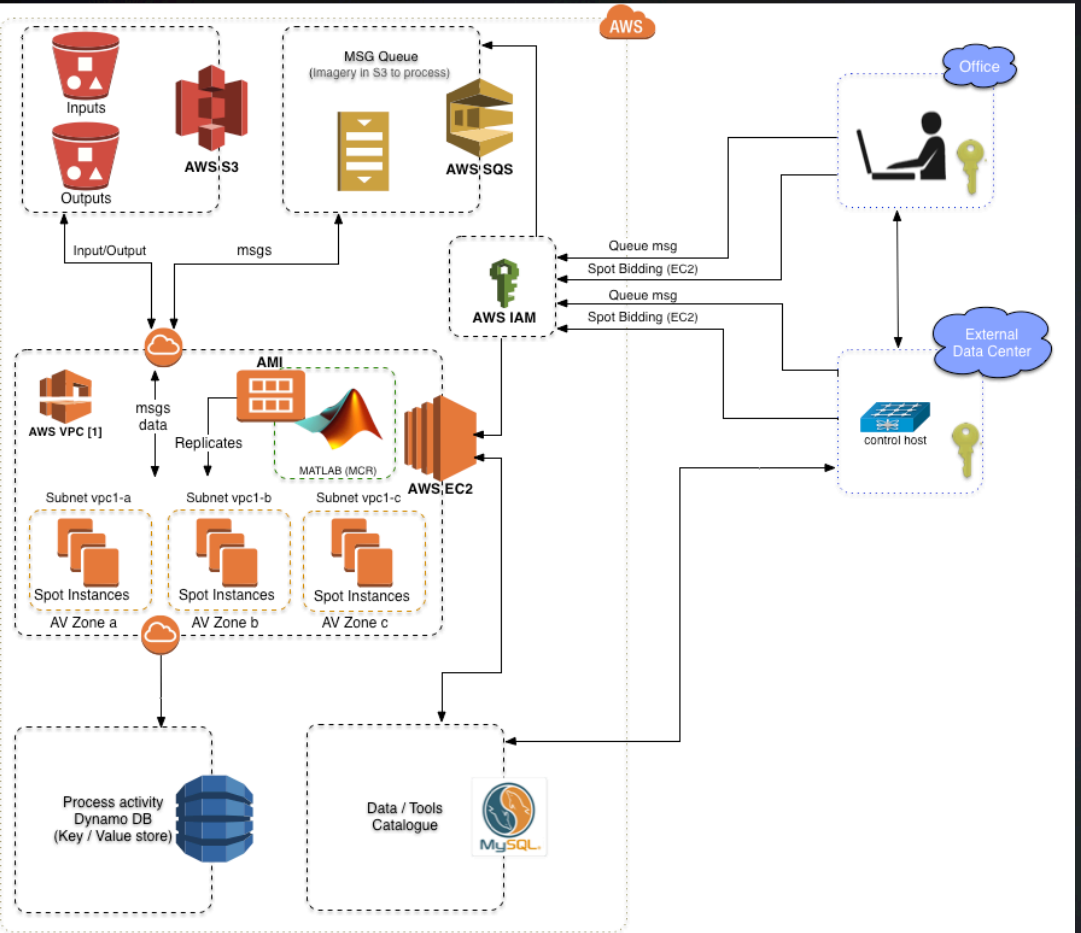
HPC M2 **HPC Module 2**

HPC M3 **HPC Module 3**

NEX **NASA Earth Exchange Storage**

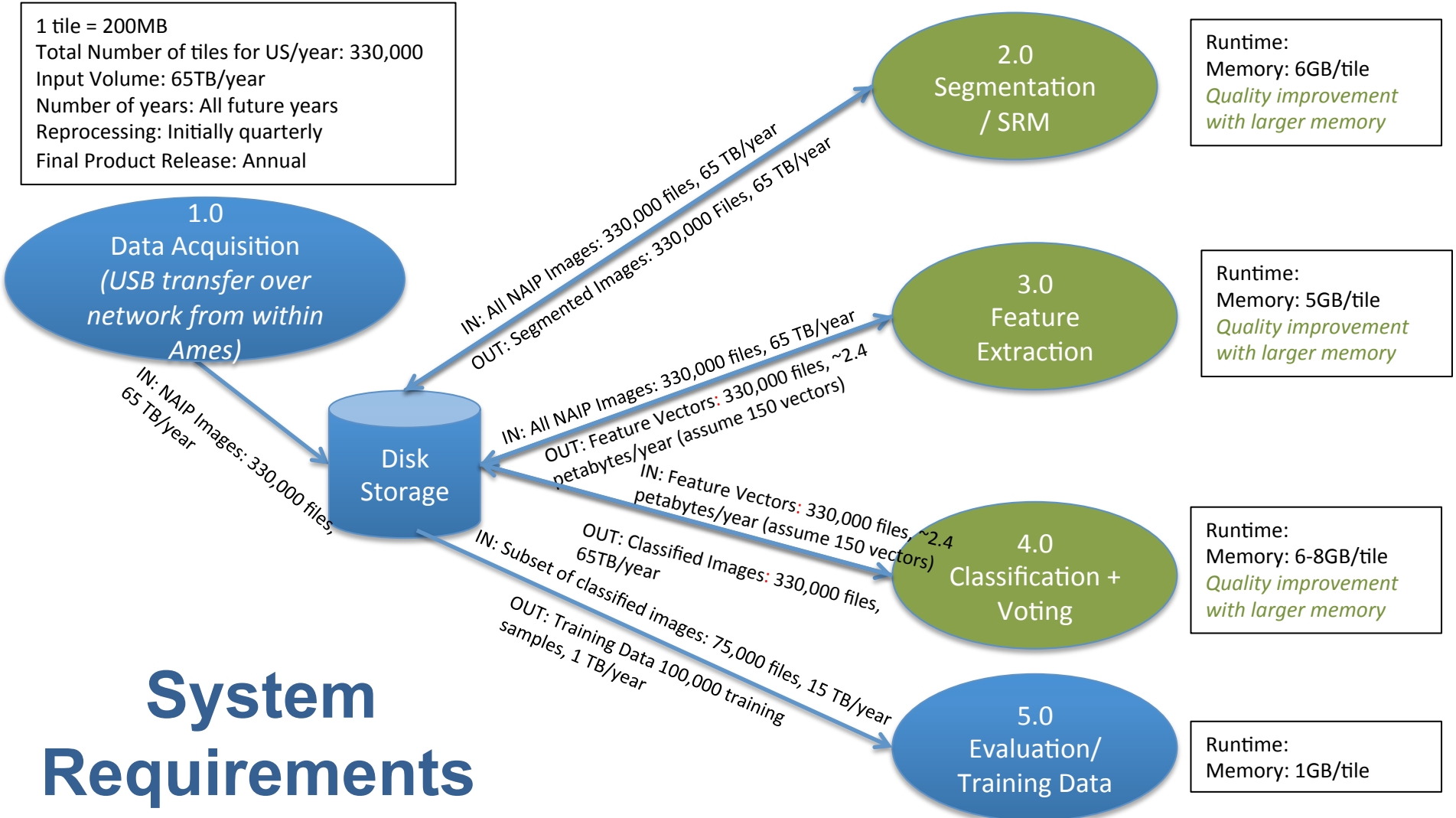
NEX HPC **NASA Earth Exchange High Performance Computing (HPC)**

National Agriculture Imagery Program (NAIP) Example



- Configure a base set of AWS services to build the processing pipeline
- Process ~15,000 Scenes
  - ~5000 x 5000 pixels / scene
- Leveraged Spot Instances
  - 70% savings
  - Managed services
    - Spinup, process, tear down in 1 week.
- More that just computing…

1 tile = 200MB
Total Number of tiles for US/year: 330,000
Input Volume: 65TB/year
Number of years: All future years
Reprocessing: Initially quarterly
Final Product Release: Annual



System Requirements

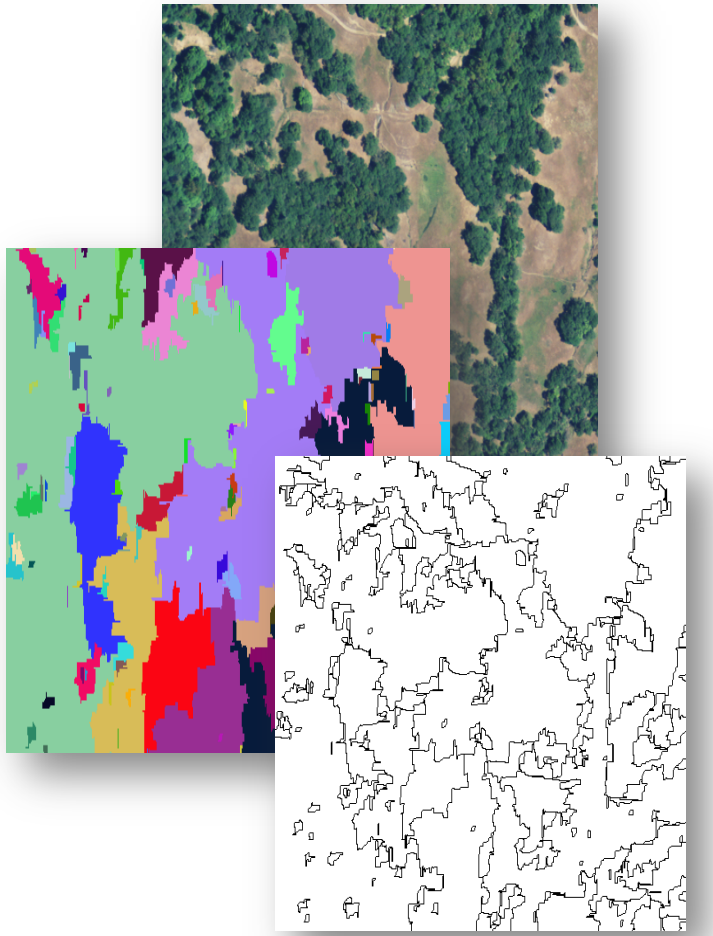
Segmentation

A segment can be considered to be any region having pixels with uniform spectral characteristics

What is a segment?

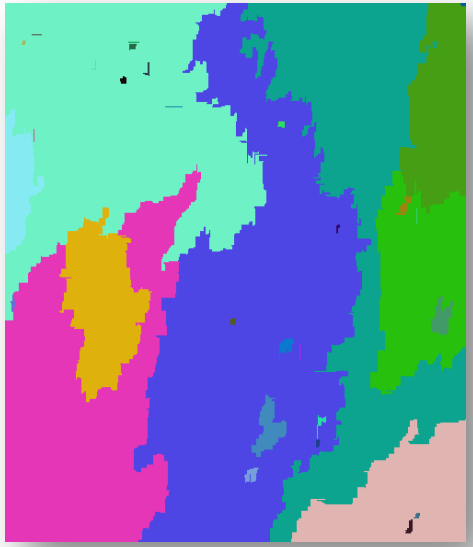
To cluster together similar looking image patches

The goal of segmentation



Segmentation using SRM algorithm

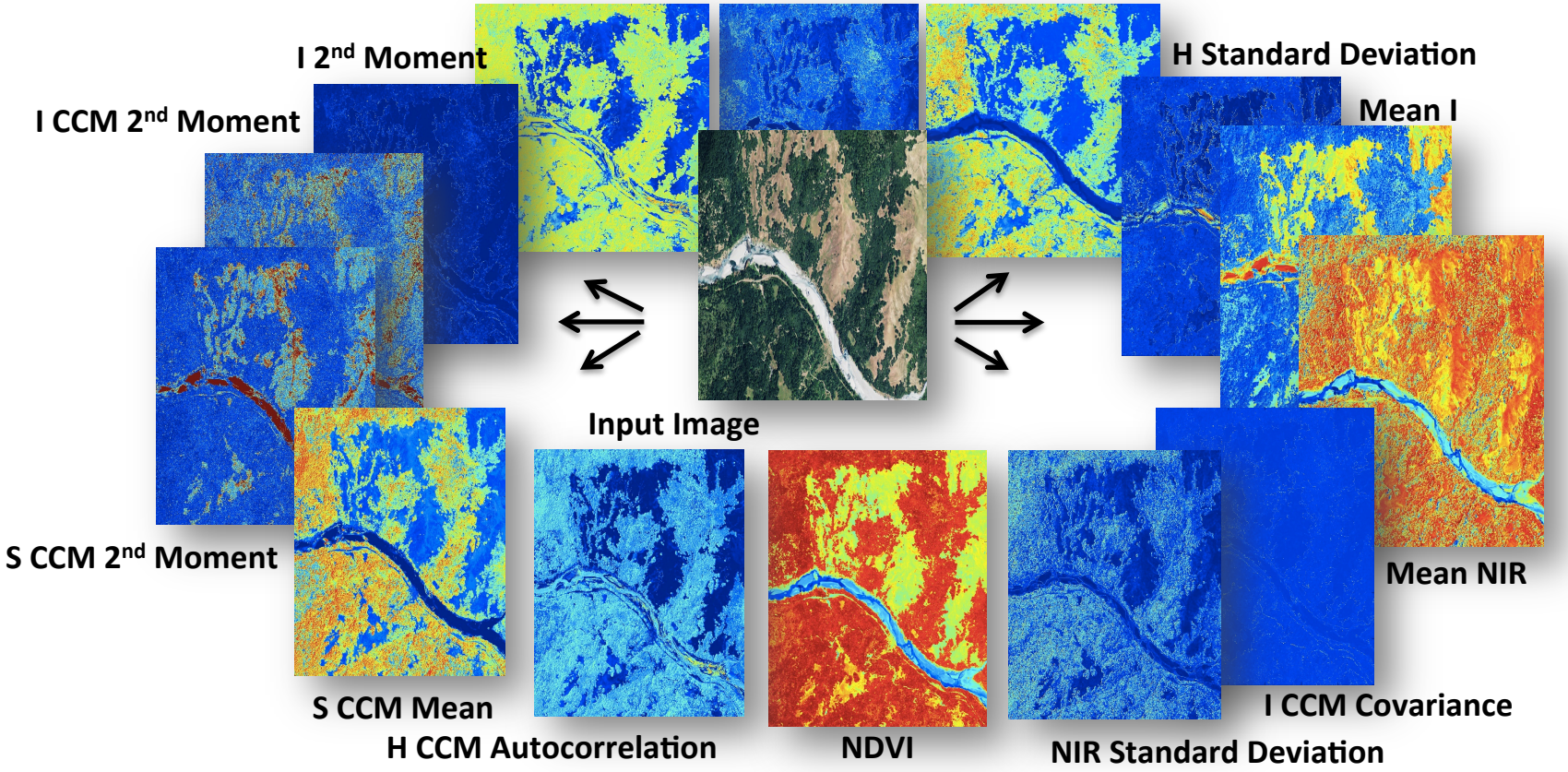
Input Image



Under-segmentation
Creates inter-class overlap within a segment

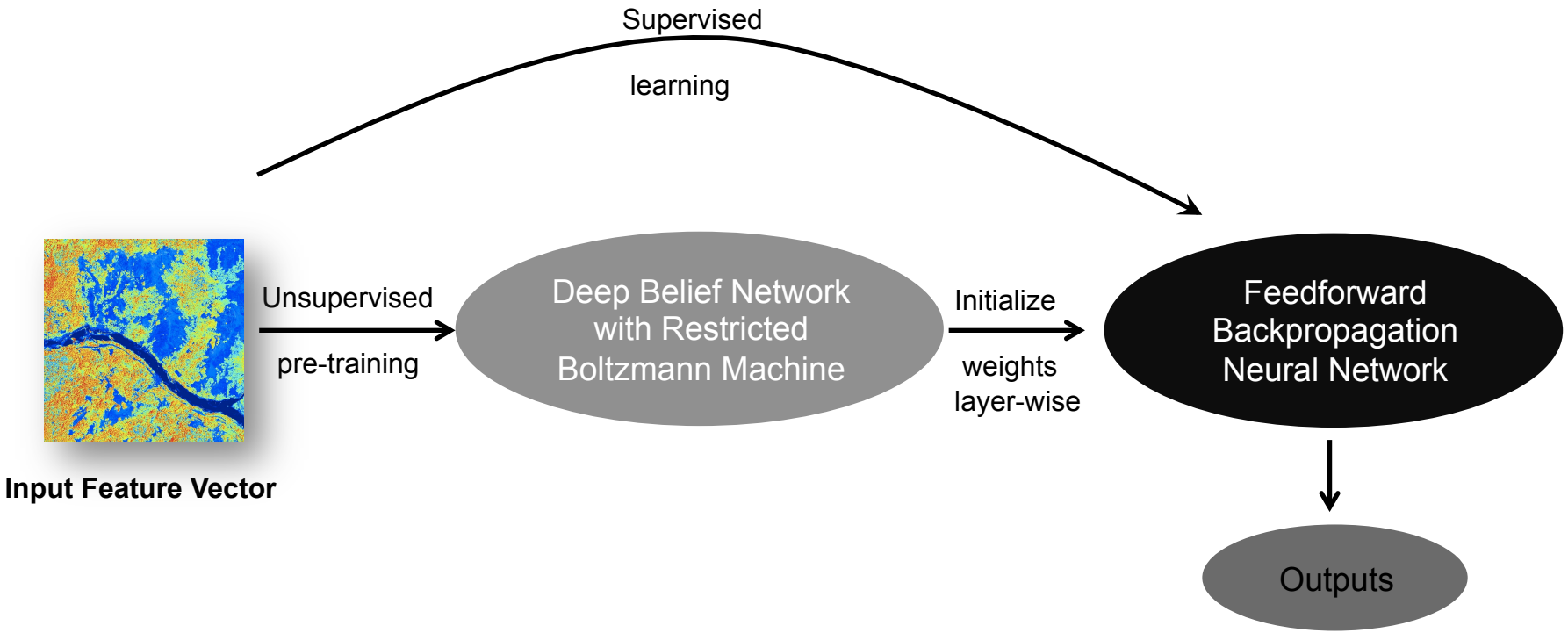
Over-segmentation
Each segment ideally contains regions belonging to a single class, no inter-class overlap

NAIP Feature Extraction Process



Multiple Features extracted from the Input Image

Learning


Supervised
learning
Unsupervised
pre-training
Deep Belief Network
with Restricted
Boltzmann Machine
Initialize
weights
layer-wise
Feedforward
Backpropagation
Neural Network
Outputs
Input Feature Vector

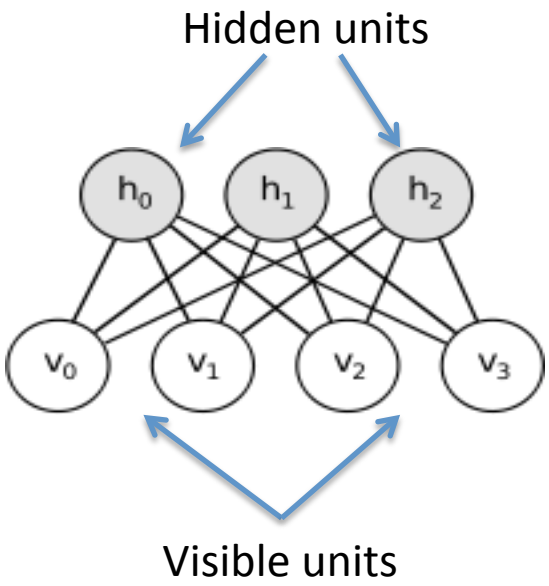
Learning

Unsupervised Learning using Deep Belief Network:

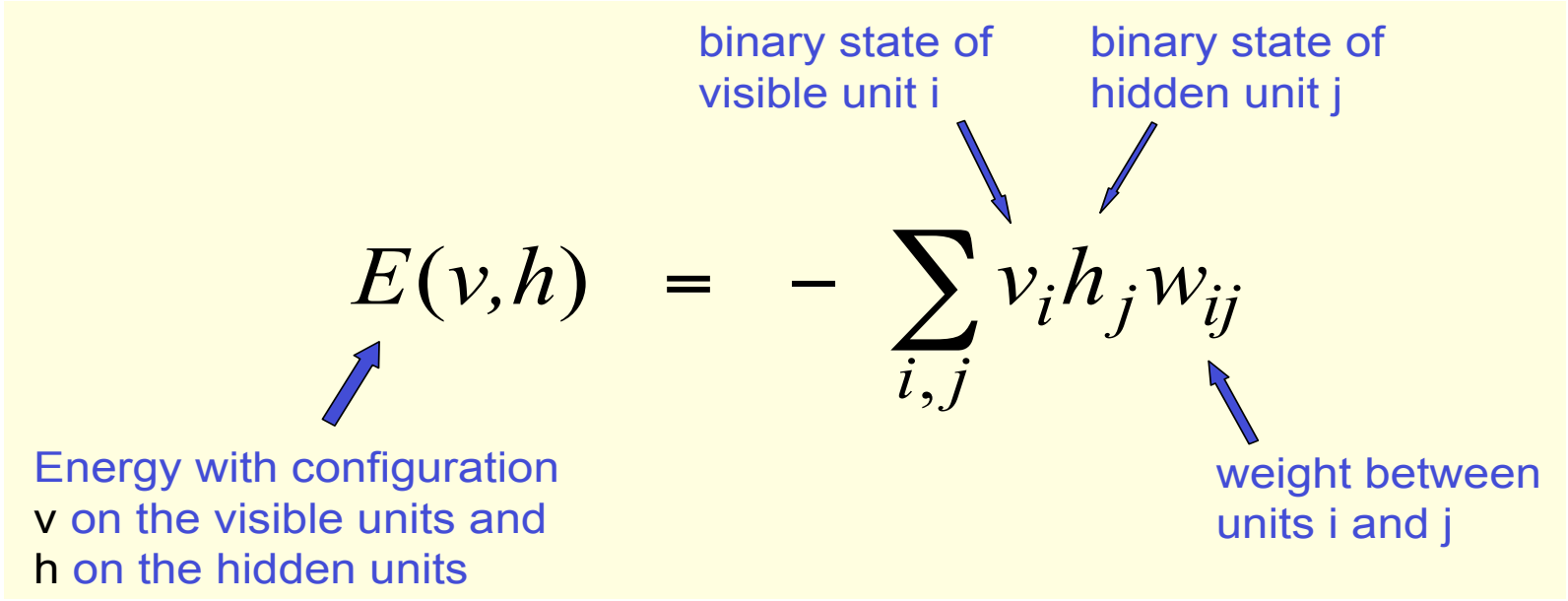
- Unsupervised pre-training using a Deep Belief Network (DBN) where each layer is trained using a Restricted Boltzmann Machine (RBM)
- The weights of the DBN are used to initialize the corresponding weights of the Neural Network
- A Neural Network initialized in this manner converges much faster than an otherwise uninitialized Neural Network
- Unsupervised pre-training is an important step in solving a prediction problem with petabytes of data with high variability

Restricted Boltzmann Machines

- Consists of 1 visible layer and 1 hidden layer.
- No connectivity between hidden units.
- Hidden units are conditionally independent given the visible states.



Energy of the Joint Configuration



Energy of the Joint Configuration (contd.)

The energy of the joint configuration gives the following gradient with respect to the weight vector

$$-\frac{\partial E(v,h)}{\partial w_{ij}} = v_i h_j$$

Each possible joint configuration of the vectors *v* and *h* has a corresponding energy.

The energy of a joint configuration determines its probability

$$p(v,h) \propto e^{-E(v,h)}$$

The Maximum Likelihood (ML) Learning for RBM

Start with the training vector clamped to the visible units.

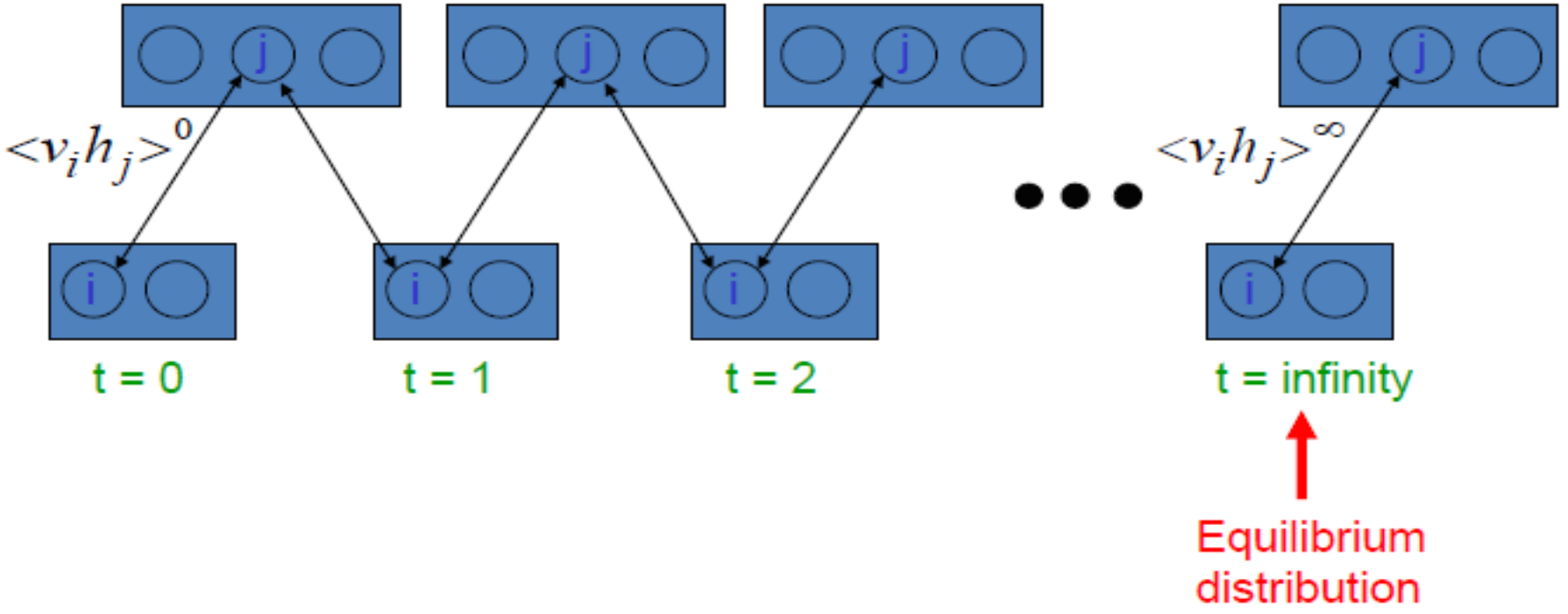
Then we alternately update all the hidden units in parallel followed by all the visible units in parallel and so on.

This continues till the model distribution reaches its equilibrium value.

The maximum likelihood learning rule can be formally written as follows:

$$\frac{\partial \log p(v)}{\partial w_{ij}} = <v_i h_j>^0 - <v_i h_j>^\infty$$

The Maximum Likelihood (ML) Learning for RBM (contd.)



The Contrastive Divergence Learning for RBM – the faster alternative to ML

Start with the training vector clamped to the visible units.

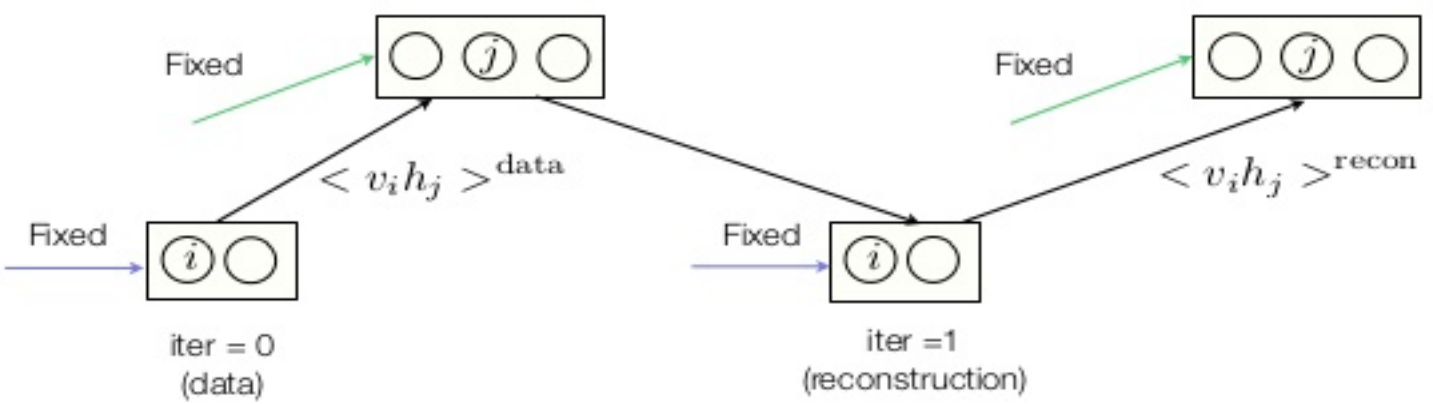
Then we update all the hidden units in parallel followed by all the visible units in parallel and finally all the hidden units.

So, unlike the maximum likelihood learning rule, the Contrastive Divergence learning algorithm runs the Gibbs sampling for just 1 step.

The Contrastive Divergence learning rule can be formally written as follows:

$$\Delta w_{ij} = \varepsilon \left(\langle v_i h_j \rangle^0 - \langle v_i h_j \rangle^1 \right)$$

The Contrastive Divergence Learning for RBM – the faster alternative to ML



Learning

Deep Belief Network:

- Each layer is conditionally independent of the other
- DBN can be trained layer-wise by iteratively maximizing the conditional probability of the input vectors or visible vectors given the hidden vectors and a particular set of layer weights
- A DBN trained layer-wise with RBM can help in improving the variational lower bound on the probability of the training data under the composite learning model

Learning

Supervised Learning using Artificial Neural Network:

- Fully connected Feed-forward backpropagation neural network

- One input layer with 26 input neurons, three hidden layers each having 100 neurons and one output layer having one neuron.

$$\sigma(t) = tanh(t) = \frac{e^t - e^{-t}}{e^t + e^{-t}}$$

- Activation function: tansigmoid (tanhyperbolic)

Neural Network (contd.)

- Weights and biases initialized using: Deep Belief Network
- Performance function: mean squared error (mse)

Training:

- In the training phase around 100,000 training samples are chosen
- Chosen randomly from a multitude of scenes having various kinds of tree-cover like urban, dense, fragmented etc.

Testing:

- Testing involves using the trained model to generate classification maps for satellite images from the dataset on the fly.

Learning Module

Training Phase

Training data

EXTRACT FEATURE VECTORS

CCM
DCT
NDVI
EVI
• • •

TAKE SUB-SAMPLE OF THE FEATURE VECTORS

TRAINING CLASS LABELS

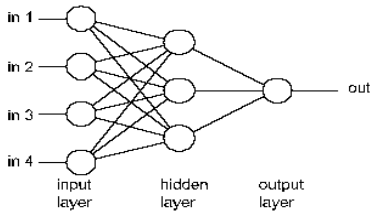
APPEND CLASS LABEL AND FEED TO ANN

UNSUPERVISED LEARNING (USING DEEP BELIEF NETWORK)

INITIALIZE WEIGHTS OF NEURAL NETWORK USING DEEP BELIEF NETWORK

SUPERVISED LEARNING (USING ARTIFICIAL NEURAL NETWORK)

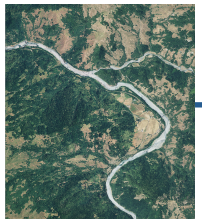
TRAIN ANN WITH BACKPROPAGATION



Trained Neural Network

Learning Module

Testing/Prediction Phase

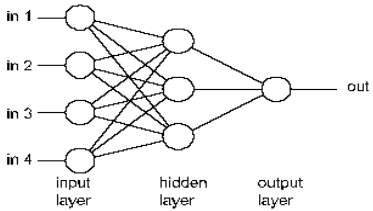


NAIP Tile

EXTRACT FEATURE VECTORS

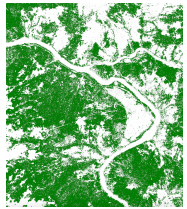
CCM
DCT
NDVI
EVI
•
•
•

NORMALIZE DATA AND FEED TO ANN



Trained Neural Network

PREDICT CLASS AND GENERATE LABELS



CLASS MASK

Experimental Results

Total scenes processed = 11095 for the whole of California

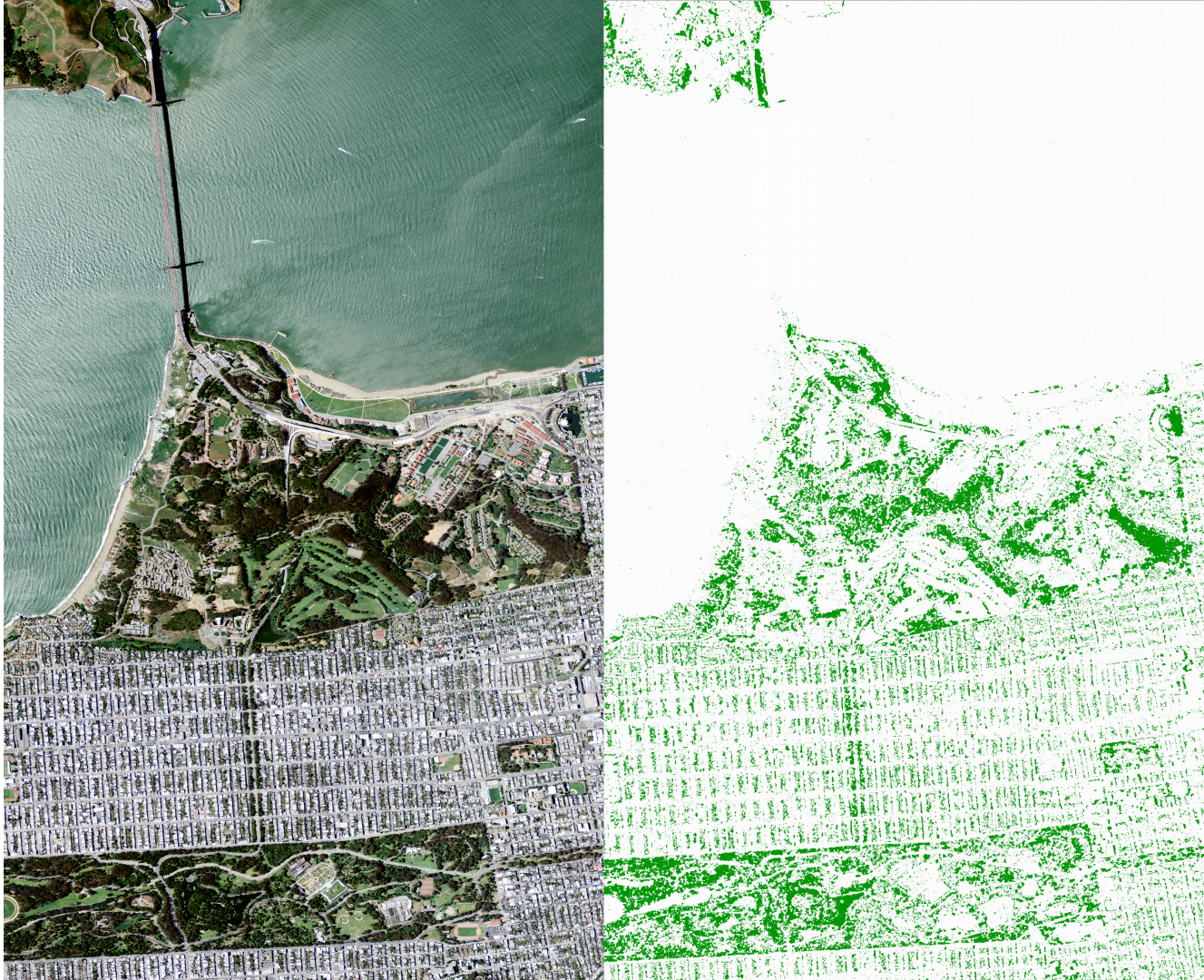
	Densely Forested	Fragmented forests	Urban areas	Overall
Total samples	12000	12000	12000	36000
Tree samples	6000	6000	6000	18000
Non-tree samples	6000	6000	6000	18000
True Positive Rate (%)	85.87	88.26	73.65	82.59
False positive Rate (%)	2.21	0.99	1.98	1.73

Comparison with National Land Cover Data (NLCD) Algorithm

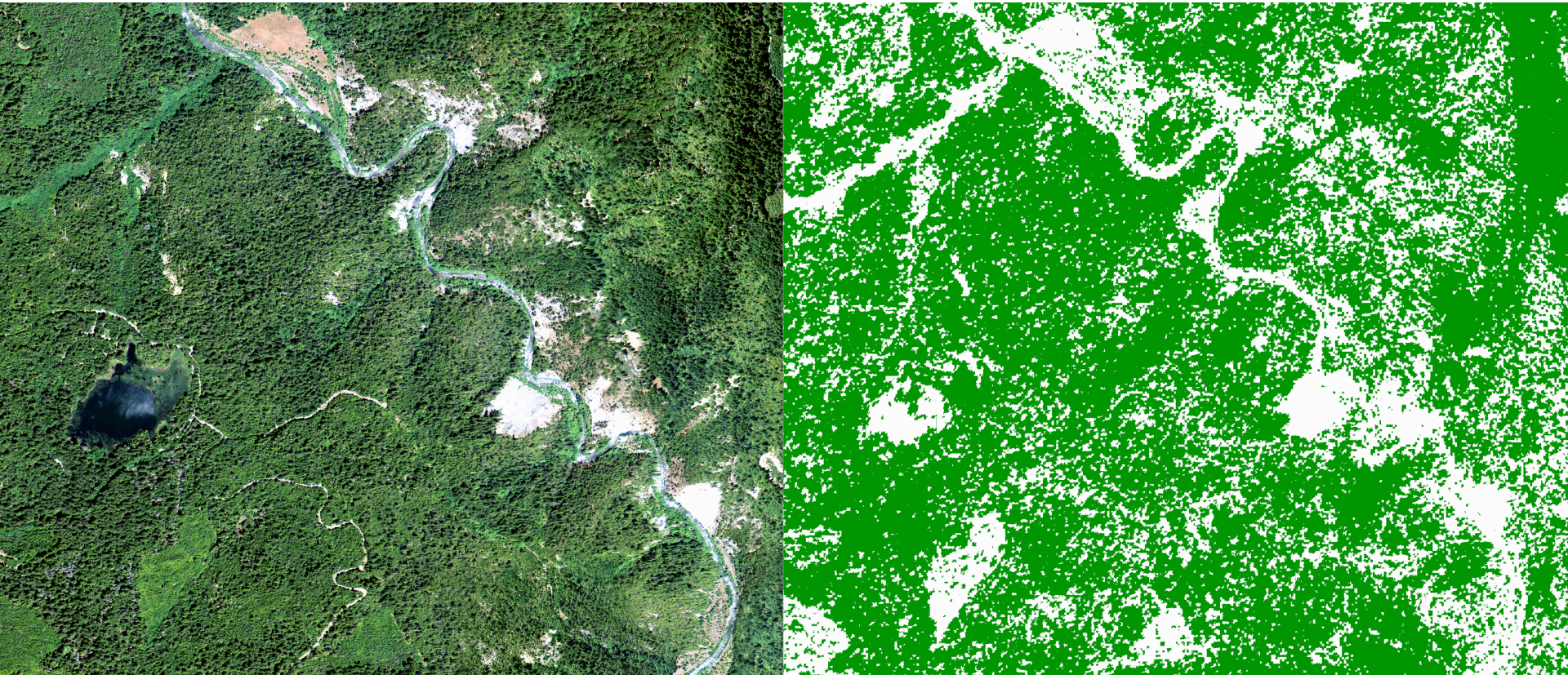
Fragmented Forests:

	NLCD 30-m	NAIP 1-m
Total samples	1000	1000
Tree samples	500	500
Non-tree samples	500	500
True Positive Rate (%)	72.31	87.13
False positive Rate (%)	50.8	1.9

San Francisco Bay Area



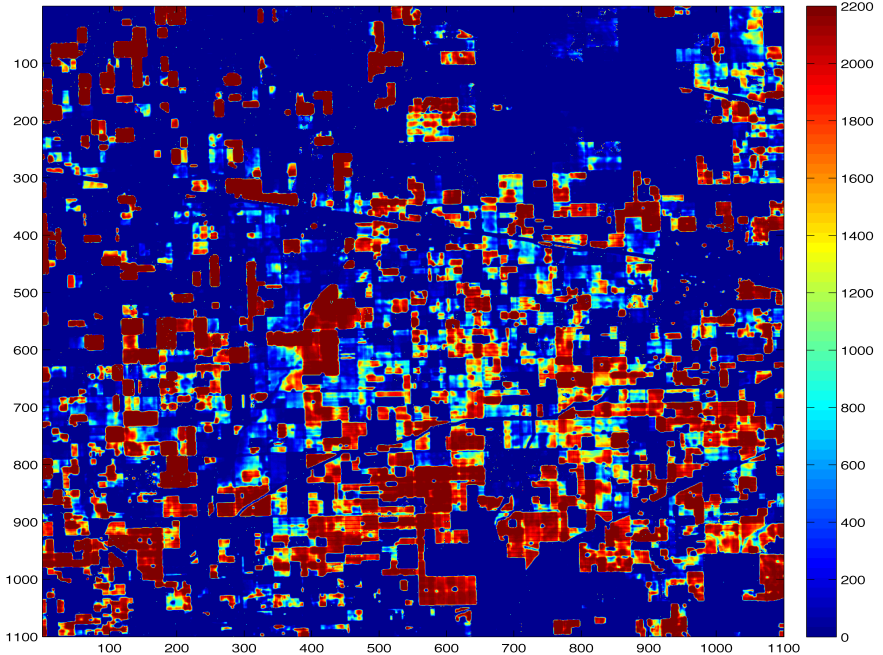
Yosemite



Yield Prediction using Deep Belief Network

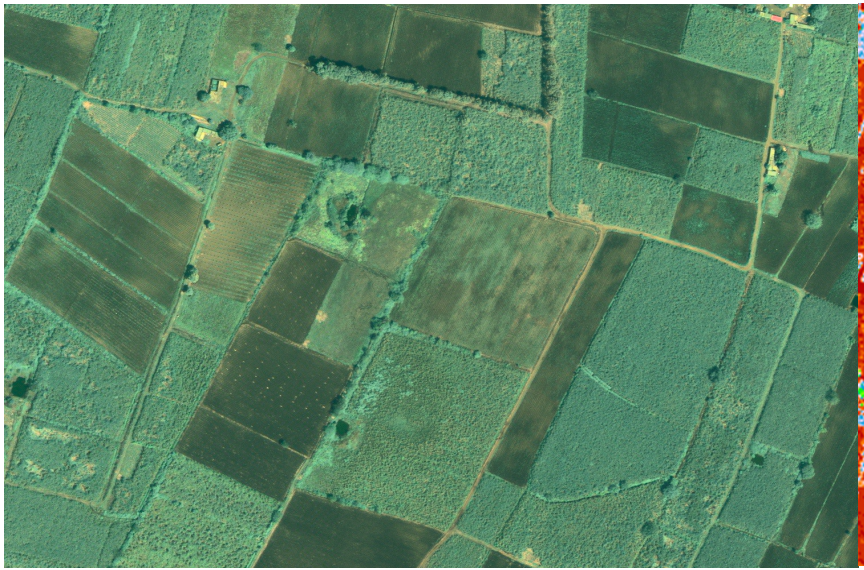
A Sample Yield Map

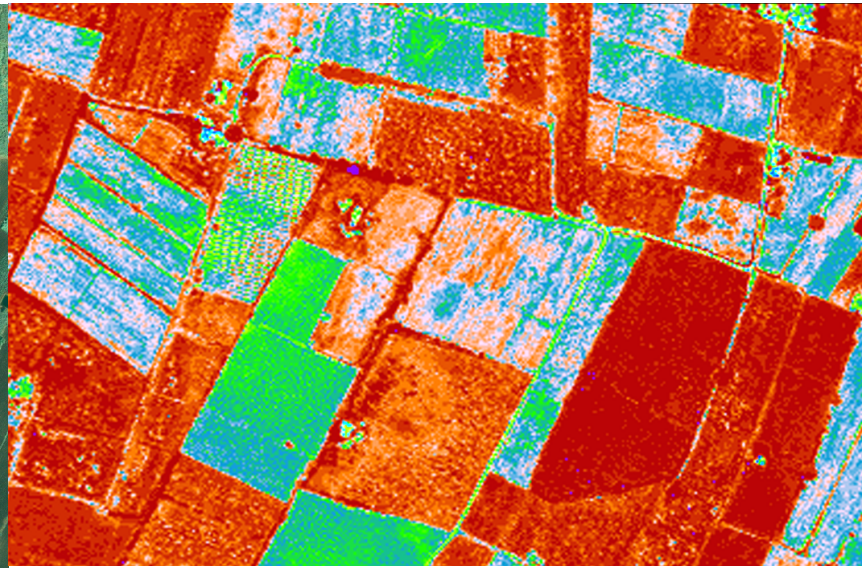
Part of a LANDSAT tile



The resulting Yield Map

Field Level Yield Prediction from High Res Imagery (WV-2)





Challenges for Yield Prediction

- Labeled training yield data is limited.
- The sample mean and standard deviations of the collected training data samples are often quite different from that of the population mean and standard deviation and hence training and test data can often represent different probability distributions.
- The amount of training data available is not enough to encode the complex higher order relationships between the various spectral bands, climate variables and the corresponding yield values.

Advantage of the Deep Belief Network based Learning Framework

- Since labeled training data is limited, we have to resort to **Unsupervised Learning**.

- **Deep Belief Networks** use unlabeled data in the first phase. Since, there are ample amounts of unlabeled data, the unsupervised learning phase is able to initialize the weights and biases of the Neural Network to a global error basin.

- Because the neural network is initialized to a global error basin, in the supervised learning phase, it requires very little training data which is well suited for our purposes since we already have limited training data.

- DBN provides the most powerful and state-of-the-art learning framework to address these problems.

SUMMARY

- NEX lowers the barrier of entry (co-locating data, model codes, and compute resources), allows knowledge sharing and provides a platform for prototyping and scaling applications

- Earth Sciences = Big Data <=> Machine Learning = Intelligent Information/ Prediction/ Estimation, etc. => Applications

- Machine Learning is great but understanding the “data” is critical.

- Scale and generalize learning models such that they are applicable to different domains – e.g. land image classification technique to ocean chlorophyll abundance and/or species classification.

THANKYOU.

FOR YOUR **ATTENTION**

exit