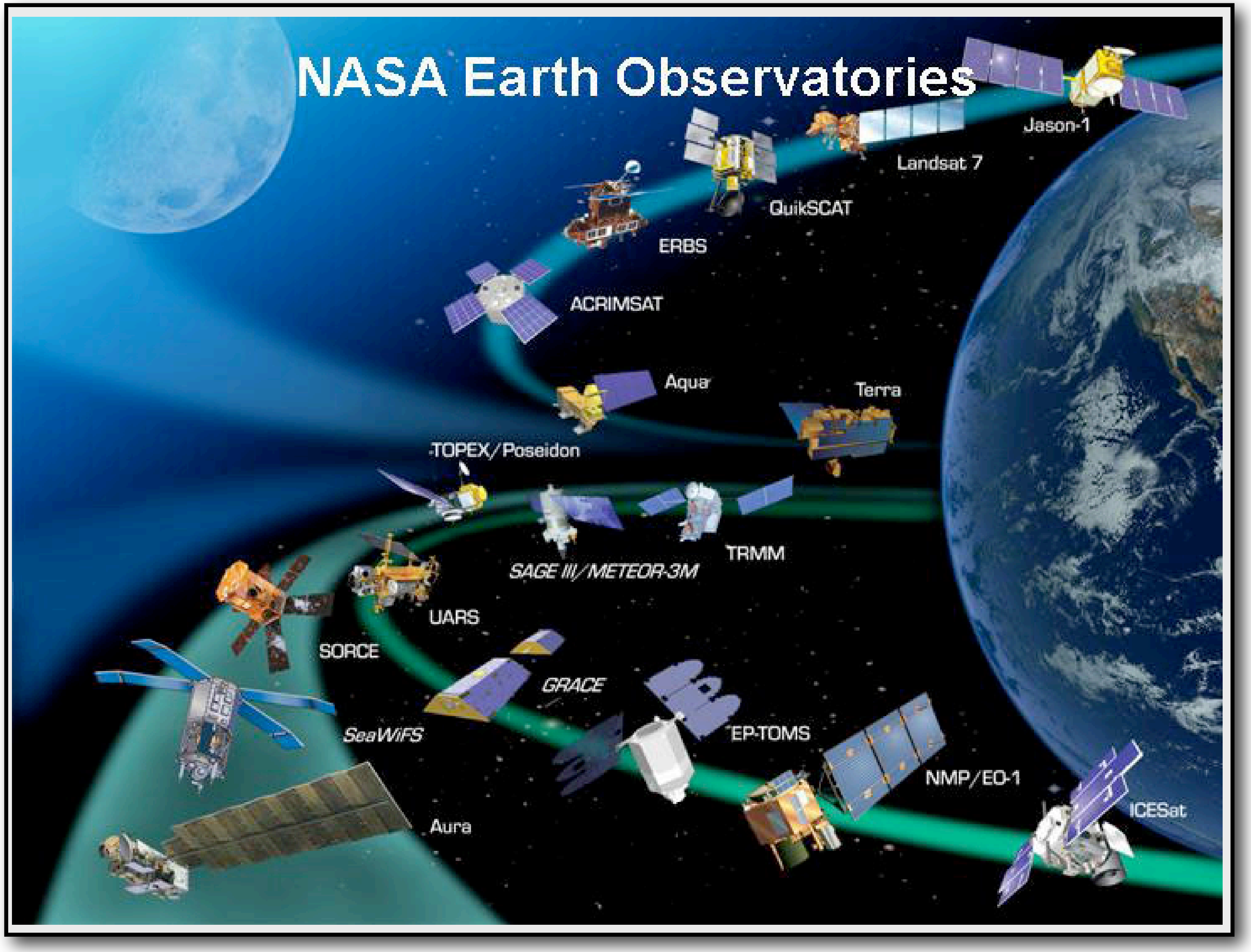
Applications of Statistical Data Analysis at CCNY and the Graphyte Toolkit

Irina Gladkova
Michael Grossberg
Dept. of Computer Science, CCNY, CUNY
NOAA/CREST

Flood of Data

50 > Multi-sensor Platforms



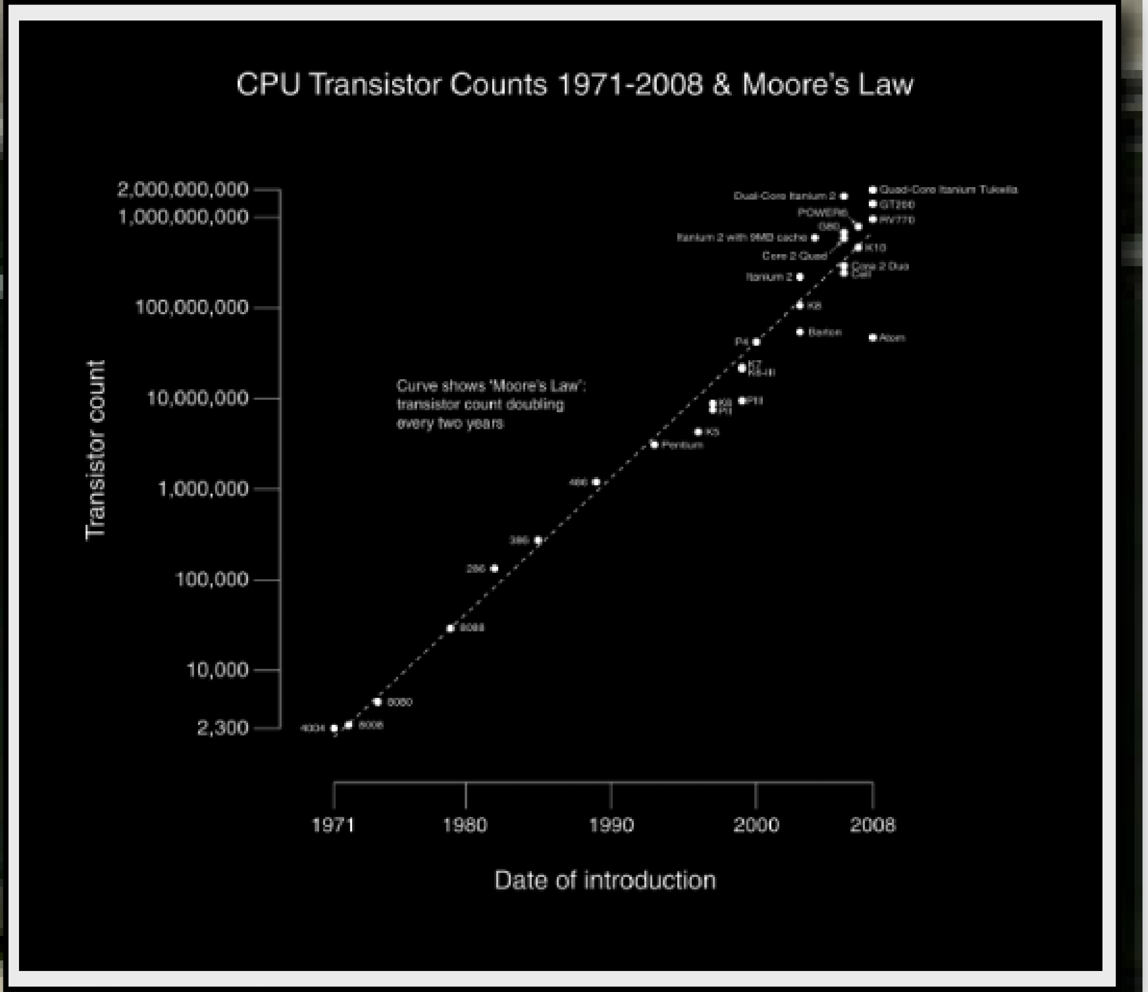
1 Sensor (MODIS) = 125 GB/DAY

GOES 9,10,12
NOAA-15,16,17,18
LandSat 5,7
DMSP F13,14,15,16
Meteosat 6,7,8,9
CBERS-2,2B
SPOT-2,4,5
ENVISAT
Resourcesat 1
CARTOSAT-1,2,2A
RADARSAT-1,2
KOMPSAT-1
THEO-1
GOMs
GMS-5
METEOR-3
OKEAN
Feng-Yun

Moore’s Law

NOAA High Performance Computing Systems

WJet System - 3376 CPUs



CPU Transistor Counts 1971-2008 & Moore’s Law

Complex Relationships

- High Dimensionality:
  - Hyper-spectral images
  - High resolution
- Non-linear relationships
- Statistical Analysis:
  - Starting point for physical modeling
  - Pre-processes for visualizations

Application and Data Driven

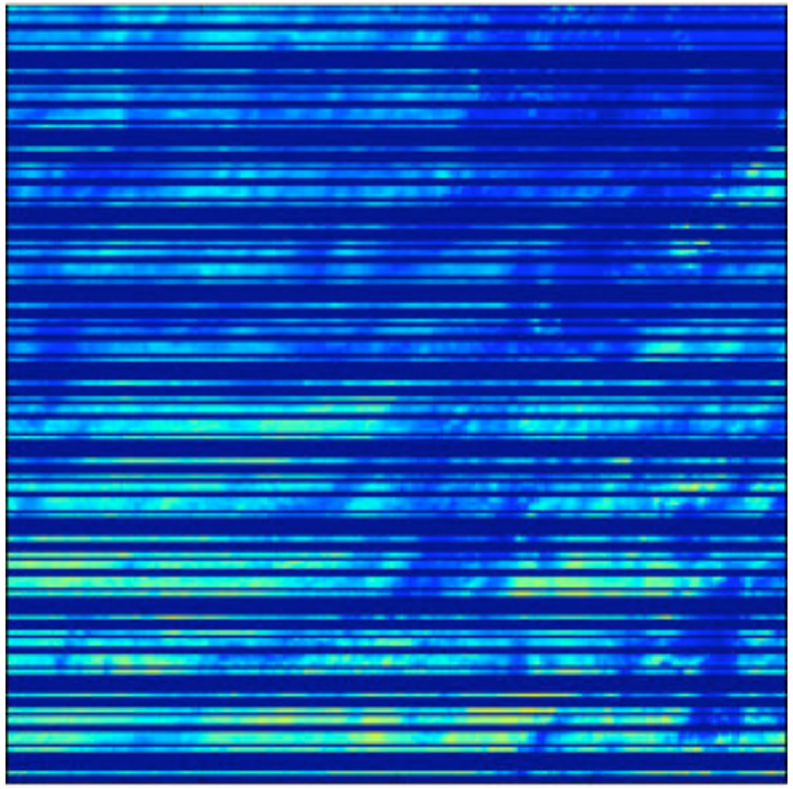
- Built tools
- Developed expertise
- Applying statistical analysis to NOAA data and problems in collaboration with NOAA Scientists

Reality: Detectors Break

Manufacturing Flaws
Launch Damage
Space is Harsh

NASA
Aqua
Earth Observing System
AIRS • AMSR-E • AMSU-A • CERES
MODIS
Moderate-resolution Imaging Spectroradiometer
NASA Goddard Space Flight Center

Band 6: 1628 - 1652 nm



Band 6: 15/20 Detectors Noisy or Totally Non-Functional

Lost Opportunity

- NASA MYD10_L2: “Aqua MODIS band 7 is used in the algorithm. The test for snow in dense vegetation in the algorithm was disabled because it was observed to result in frequent erroneous snow mapping in some situations." (http://modis-snow-ice.gsfc.nasa.gov/val.html)
- The National Snow and Ice Data Center: "Version 4 (V004) MYD29 data, the most current version available, uses Aqua/ MODIS band 7 instead of band 6." (http://www-nsidc.colorado.edu/data/myd29.html)
- NOAA/STAR: “On Aqua the retrievals are made in band 7 (2.119 µm) because of poor quality data from band 6."(Ignatov A., et al "Two MODIS Aerosol Products over Ocean on the Terra and Aqua CERES SSF Datasets")

What is ‘Plan B’?

NASA: Column-wise Interpolation?

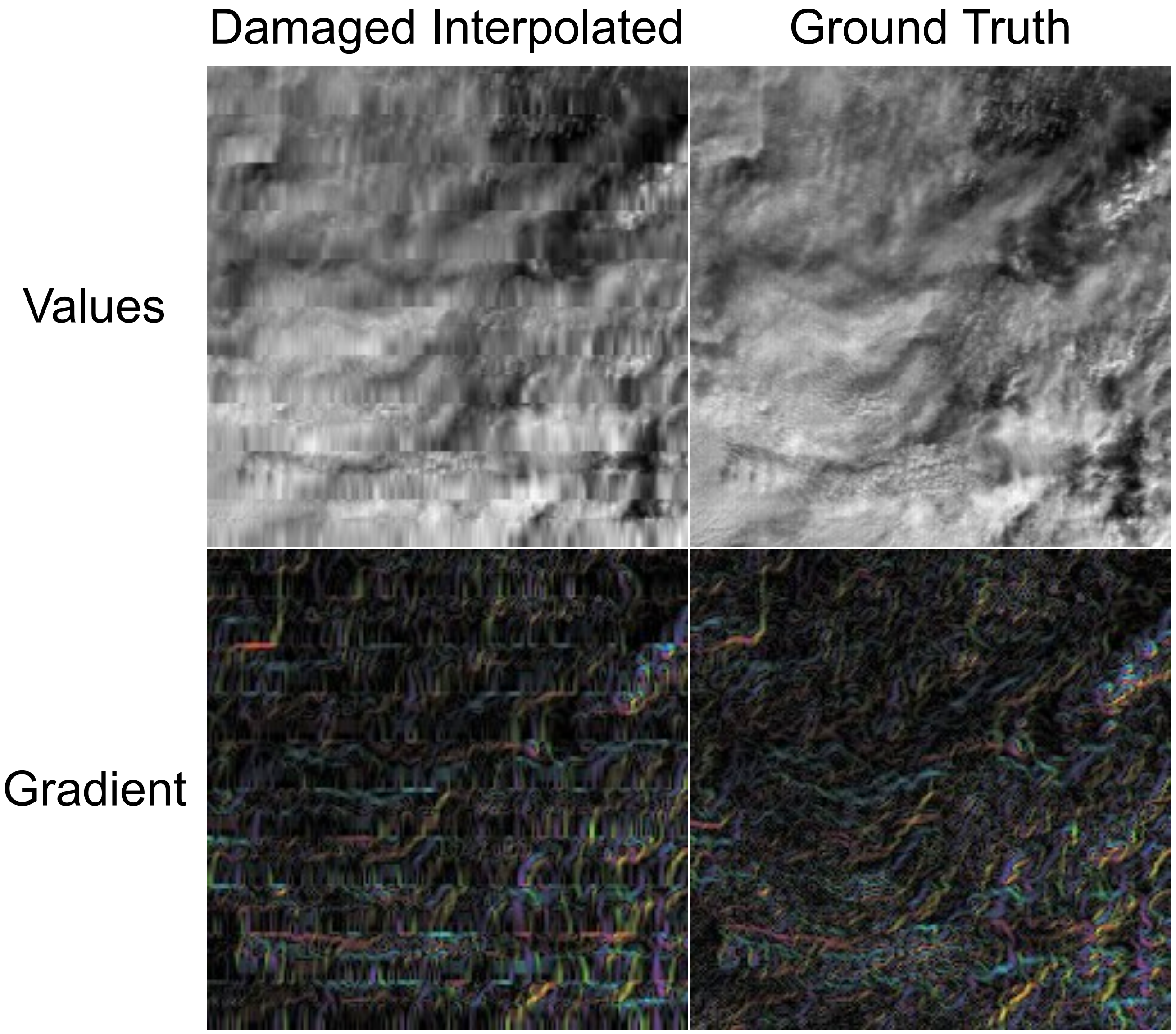
<u>Bad:</u>

Visible Artifacts

<u>Worse:</u>

Derivatives (Gradient) Fully Corrupted

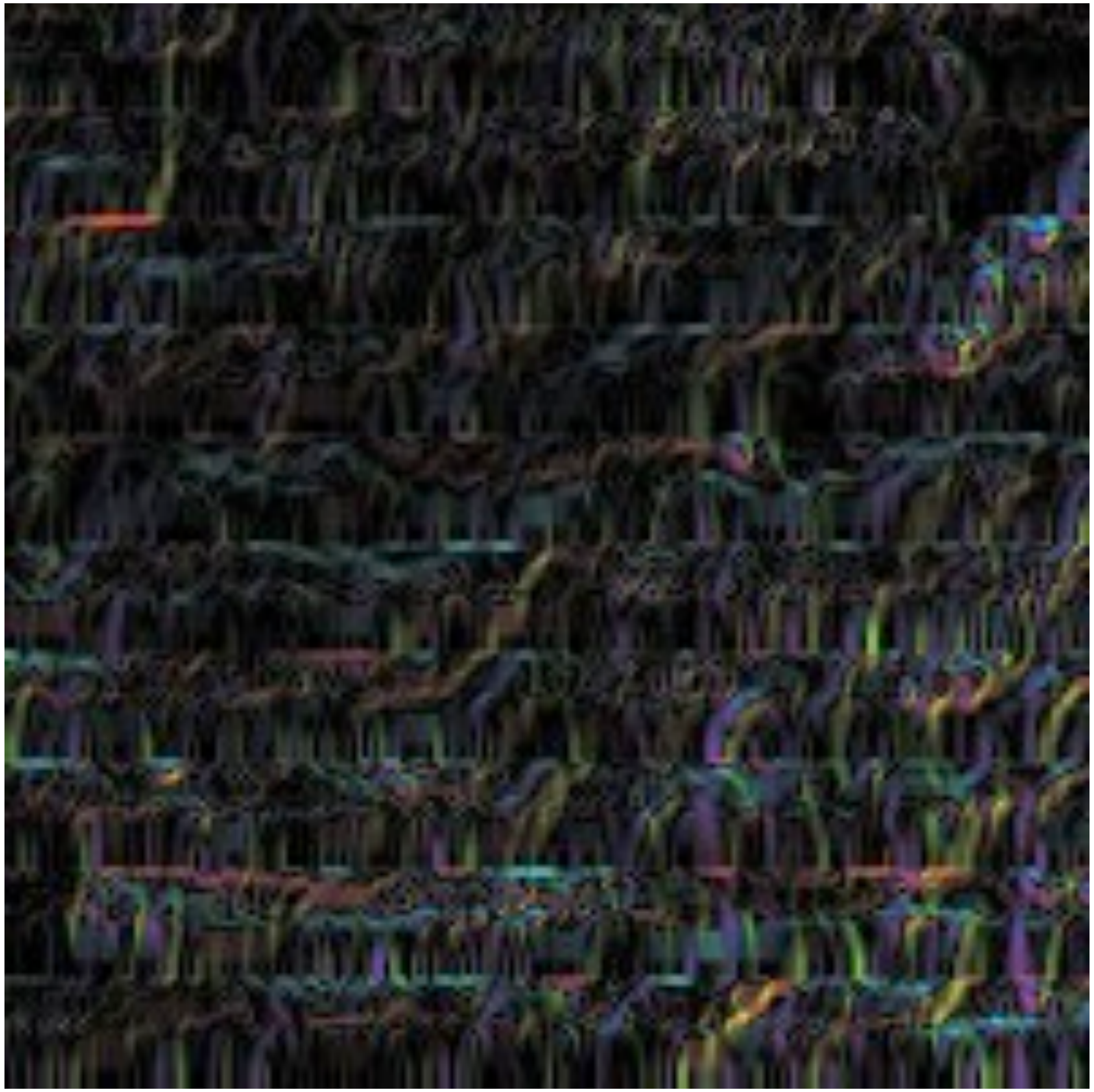
Simulate Aqua Damage with Terra for Evaluation

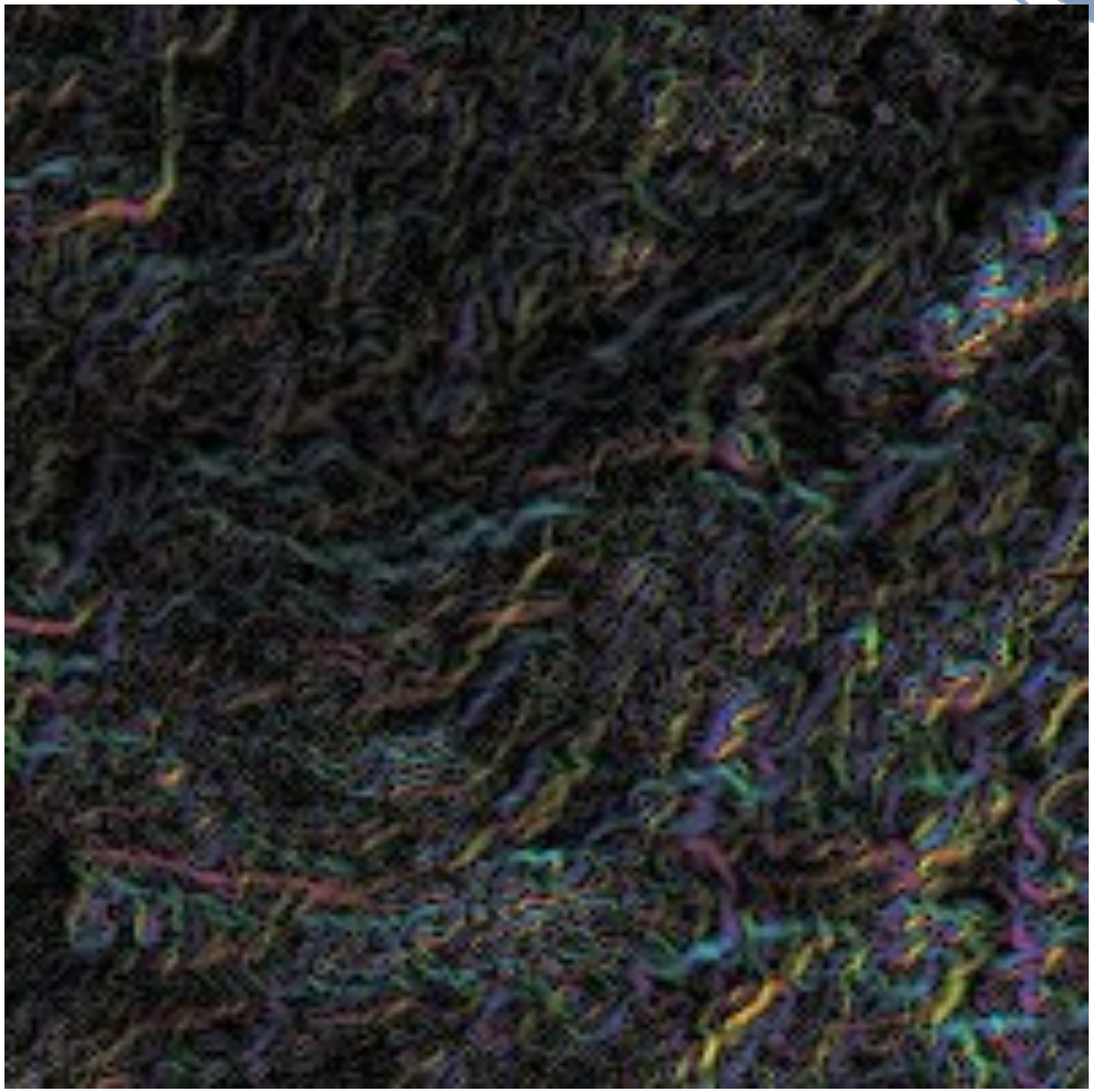


Hue = Gradient Direction, Value = Gradient Magnitude

Essential Features Destroyed

Gradients





Hue = Gradient Direction, Value = Gradient Magnitude

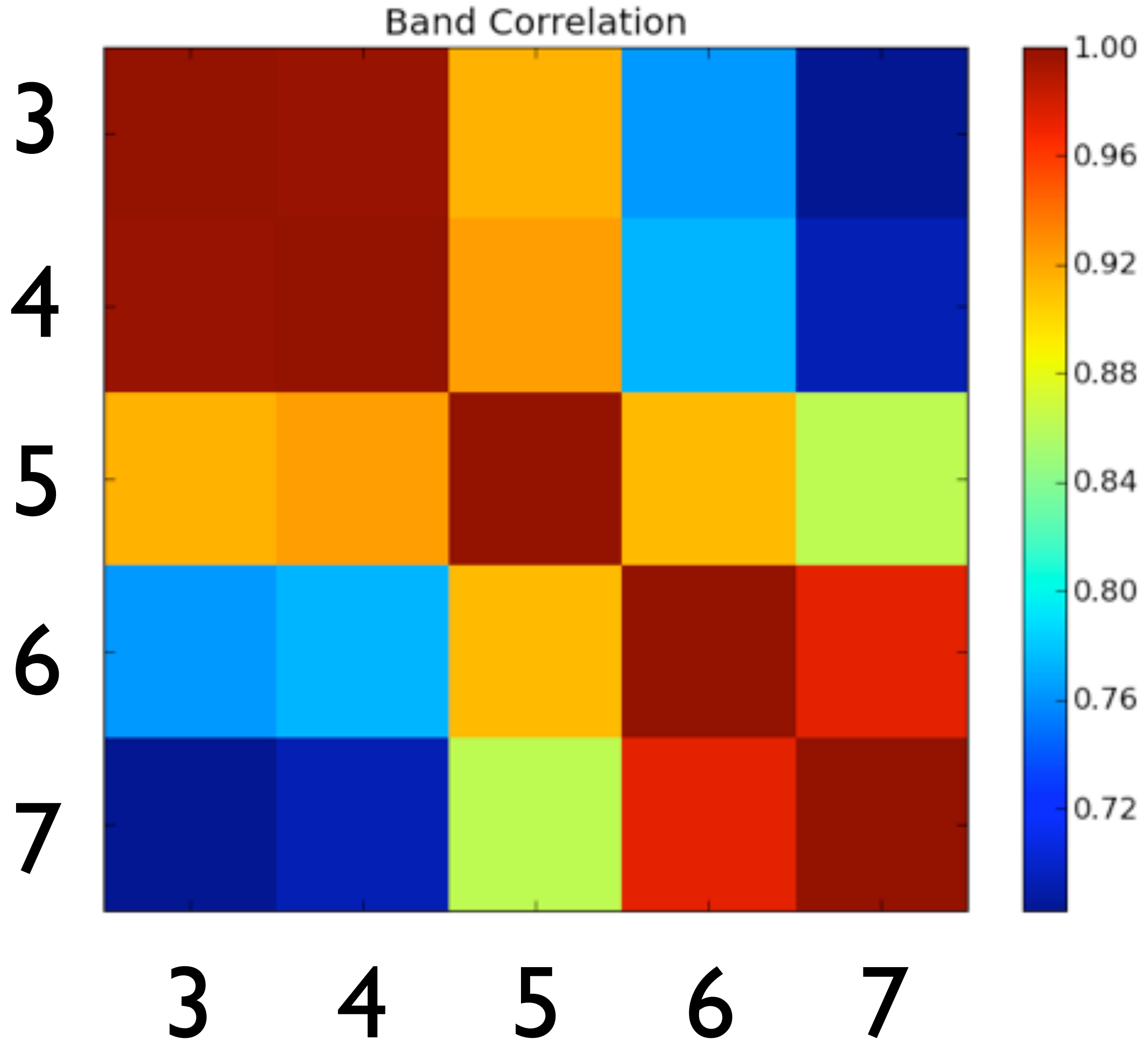
Not Much Proposed

- Only 2 papers try to fix
- Both Use Band 7 to Predict Band 6
- 2006: Global Polynomial Regression
- 2009: Local Polynomial Regression

Fundamental Problem: Band 6 not a function of 7

More Information Available

- 500m Bands have Significant Correlations
- Why not use all available information?



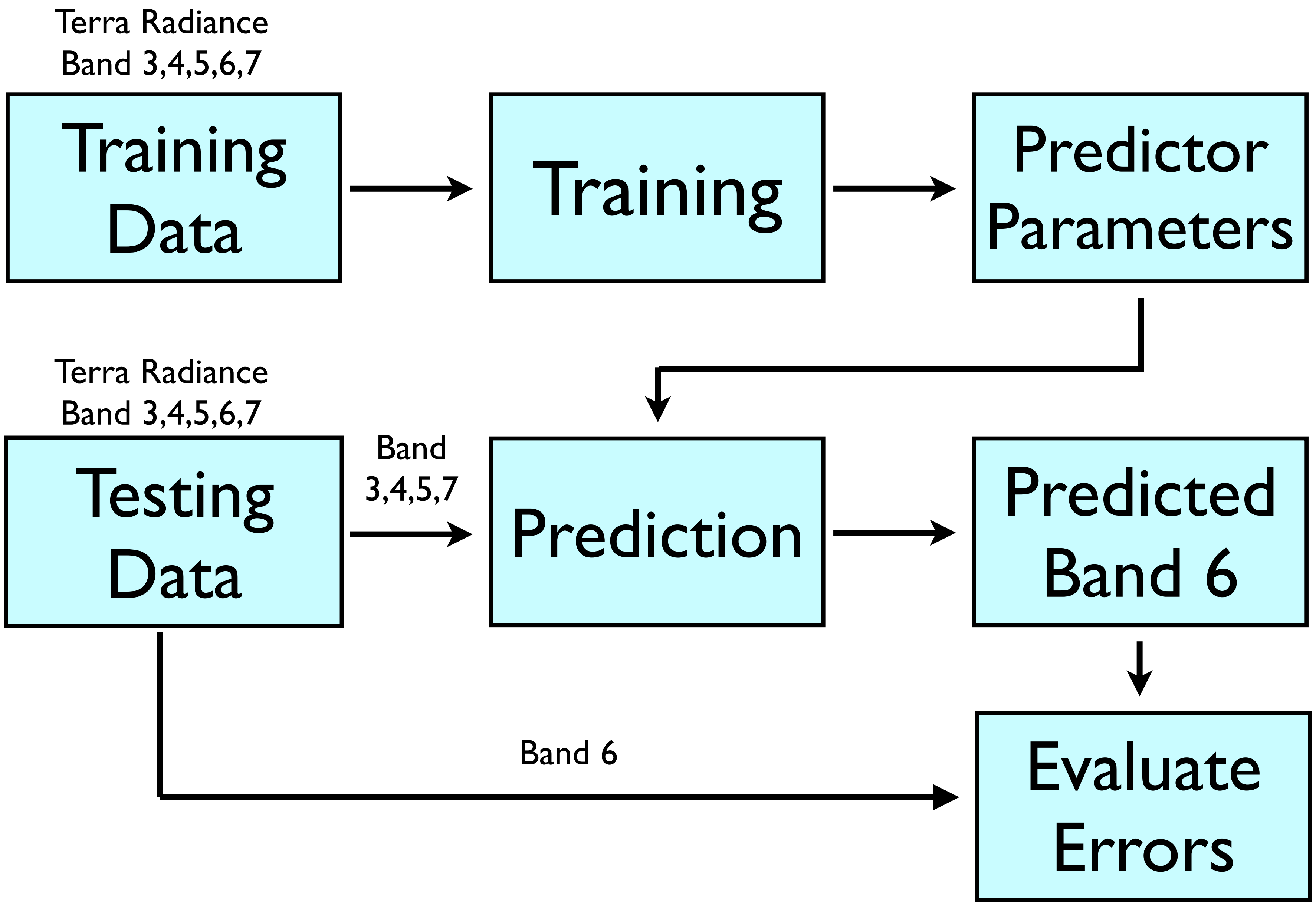
Statistical Approach

- Hypothesis:

We can predict band 6 from bands 3,4,5,7.

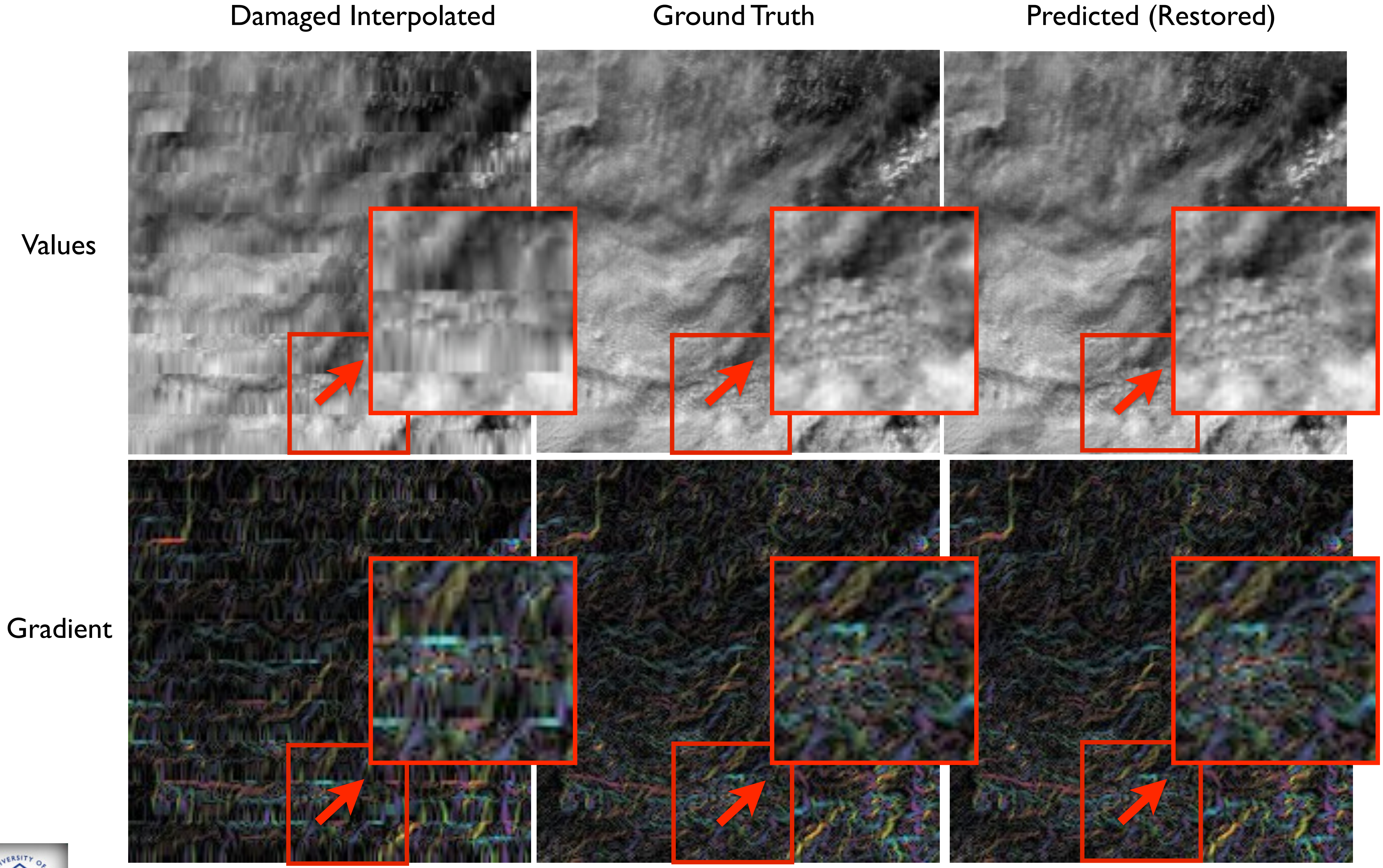
- MODIS on Terra has same bands
- Quantify prediction accuracy from test data (not used to build predictor)

Train Using Terra



Prediction used for **Quantitative** restoration

Preliminary Terra Evaluation



Damaged Interpolated

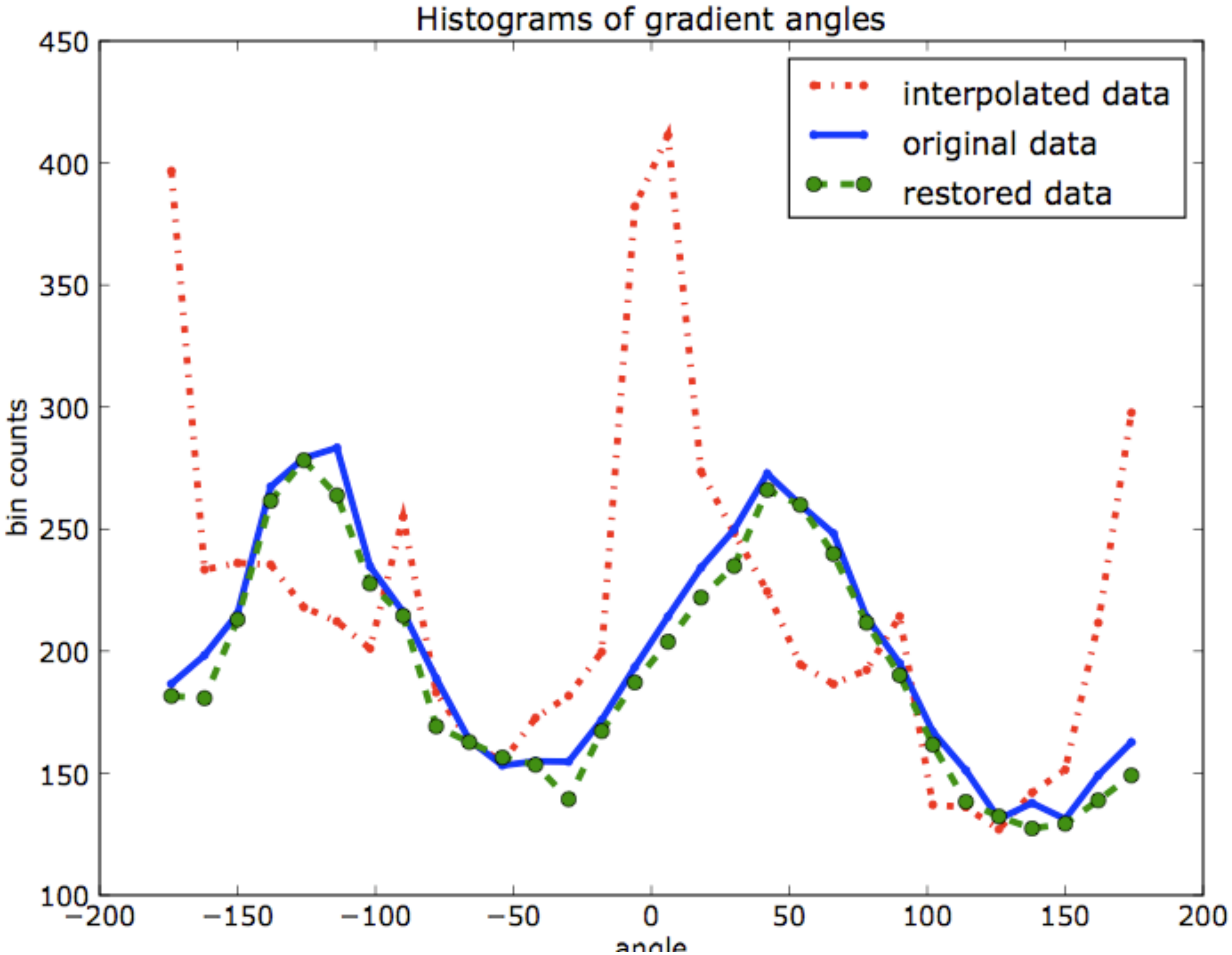
Ground Truth

Predicted (Restored)

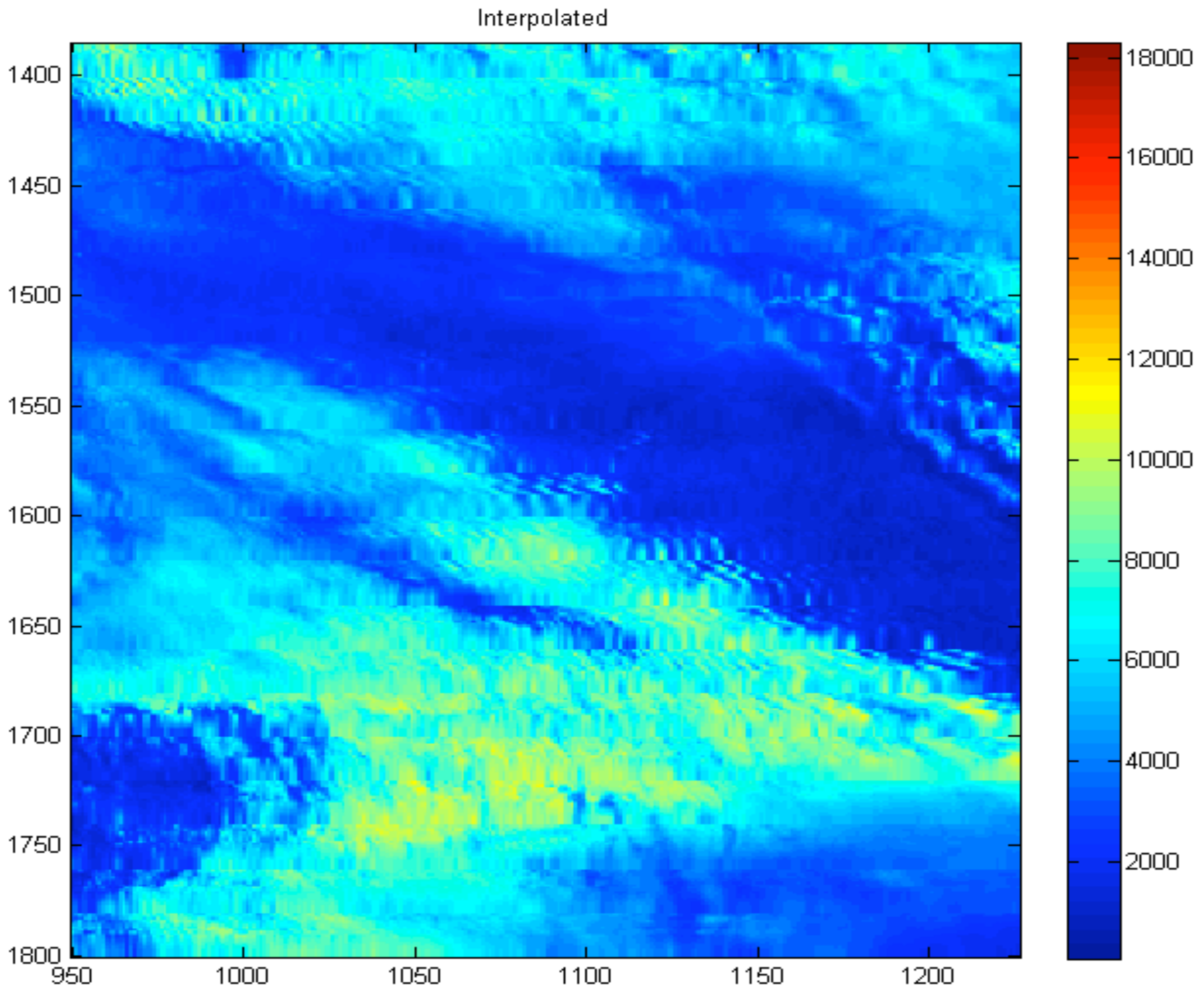
Values

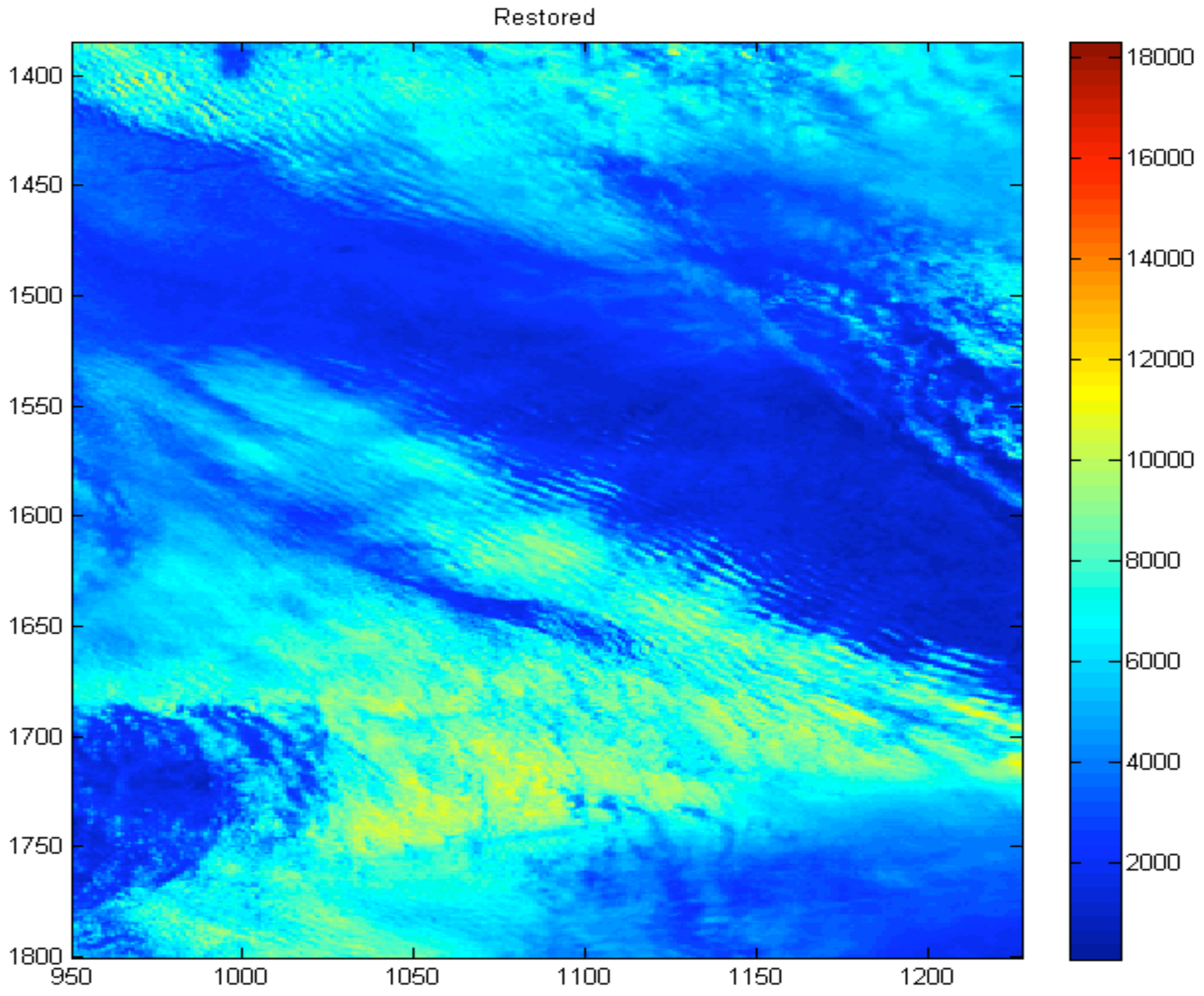
Gradient

Histogram Of Angles

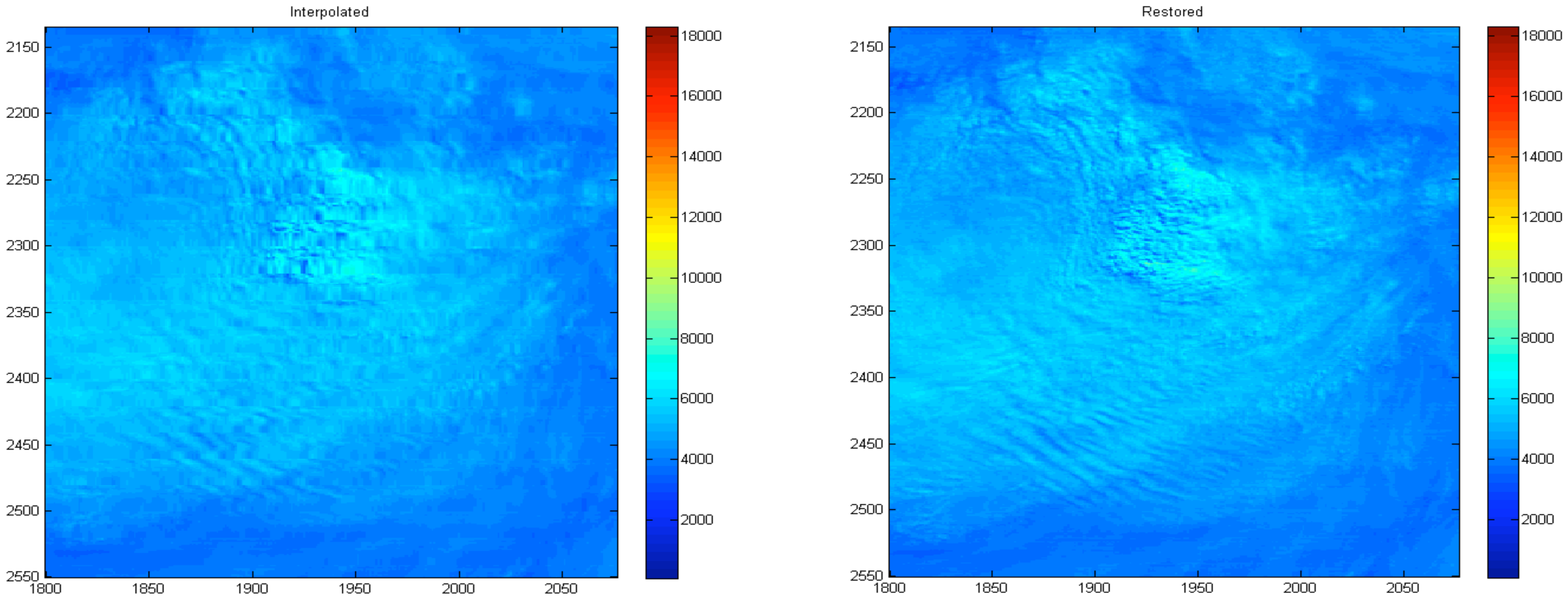


Aqua Restoration


Interpolated


Restored

Aqua Restoration



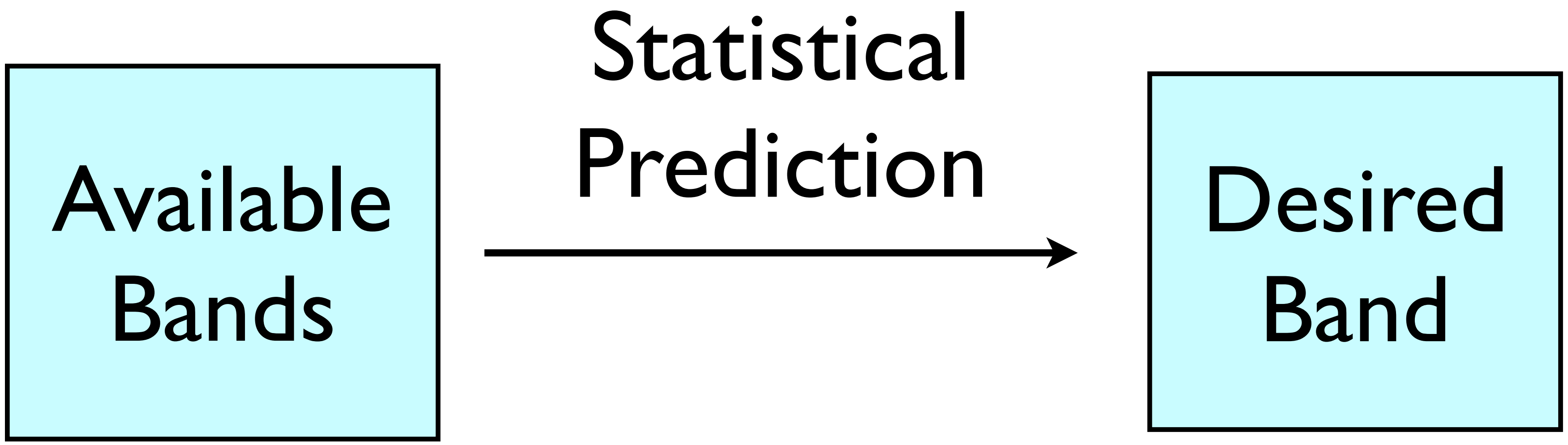
Interpolated

Restored

Evaluate For Products

- Work with STAR to potential impact for aerosol M and A products
- Investigate use for snow mapping, and cloud mask algorithms
- Adapt prediction for products directly
- Collaborate with STAR to explain physical models driving prediction

Sensor Synthesis


Available Bands

Statistical Prediction

Desired Band

Eg, Band 3,4,5,7

Eg, Band 6

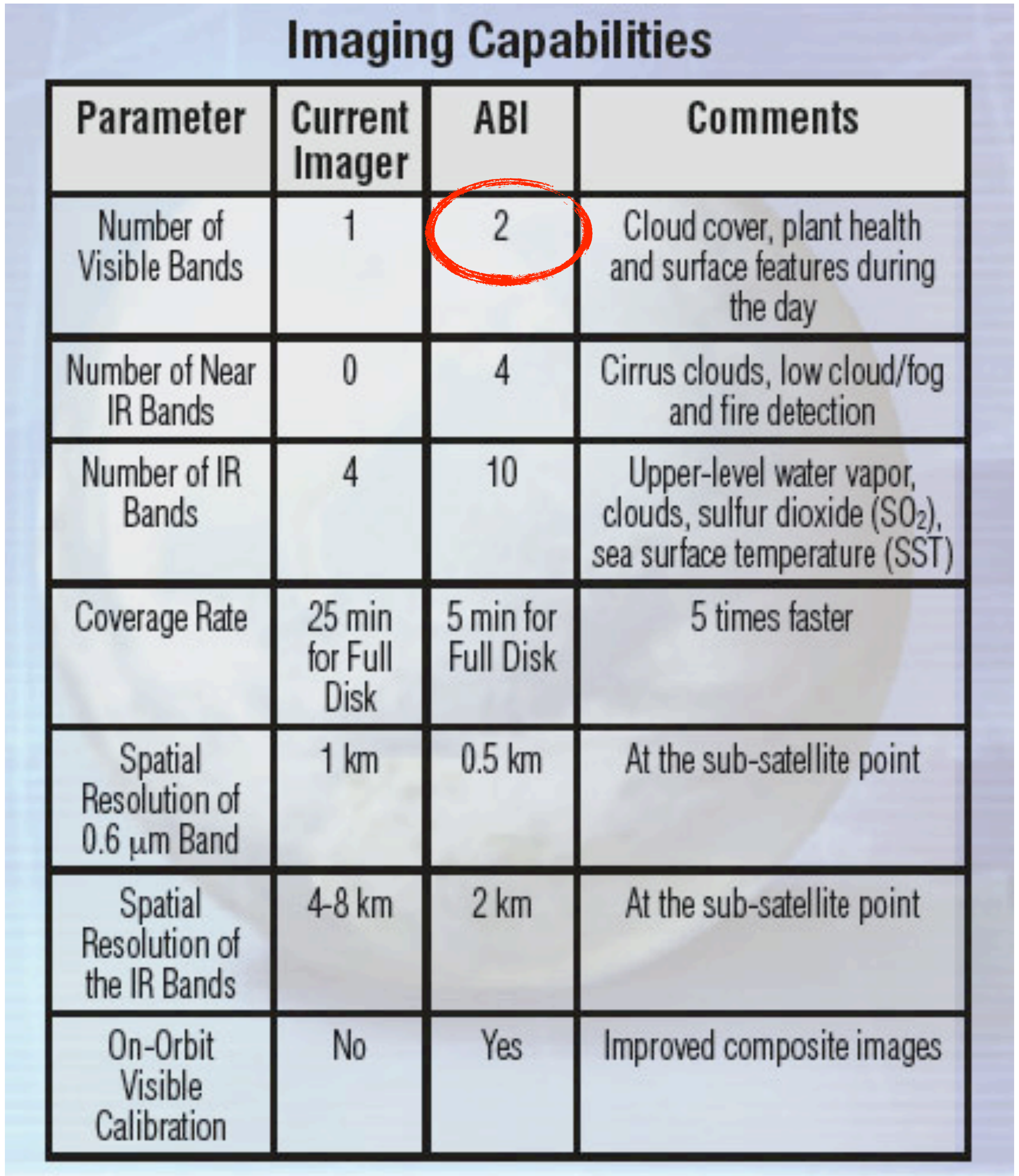
Old Elements:

- Prediction ~ Regression ~ Estimation

New Elements:

- More and higher quality data
- Much faster computers
- Able to handle non-linear multivariate problems in higher dimensions

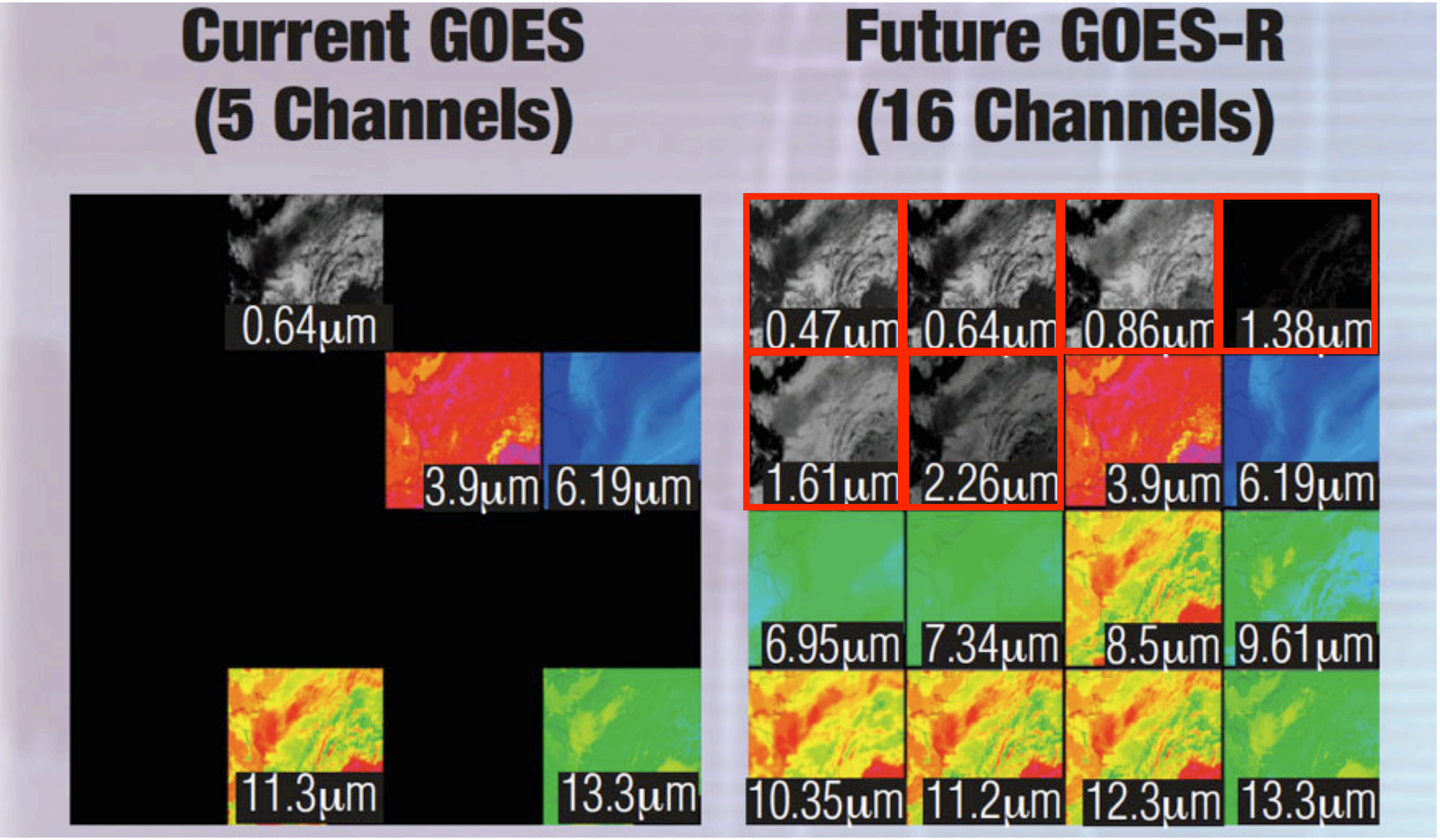
No Green on GOES-R



Imaging Capabilities

Parameter	Current Imager	ABI	Comments
Number of Visible Bands	1	2	Cloud cover, plant health and surface features during the day
Number of Near IR Bands	0	4	Cirrus clouds, low cloud/fog and fire detection
Number of IR Bands	4	10	Upper-level water vapor, clouds, sulfur dioxide (SO_2), sea surface temperature (SST)
Coverage Rate	25 min for Full Disk	5 min for Full Disk	5 times faster
Spatial Resolution of 0.6 μm Band	1 km	0.5 km	At the sub-satellite point
Spatial Resolution of the IR Bands	4-8 km	2 km	At the sub-satellite point
On-Orbit Visible Calibration	No	Yes	Improved composite images

6 Channels close to visible



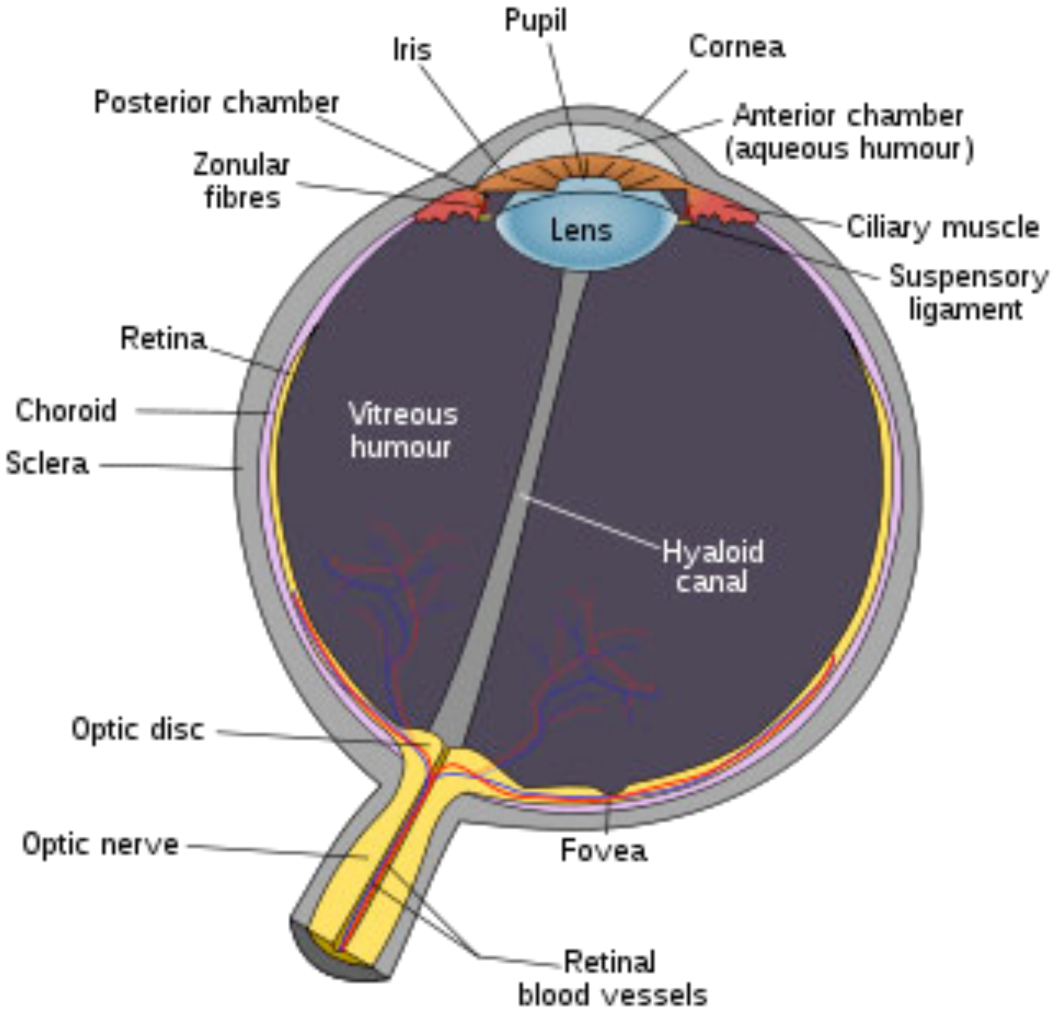
Why is Green Band Important?

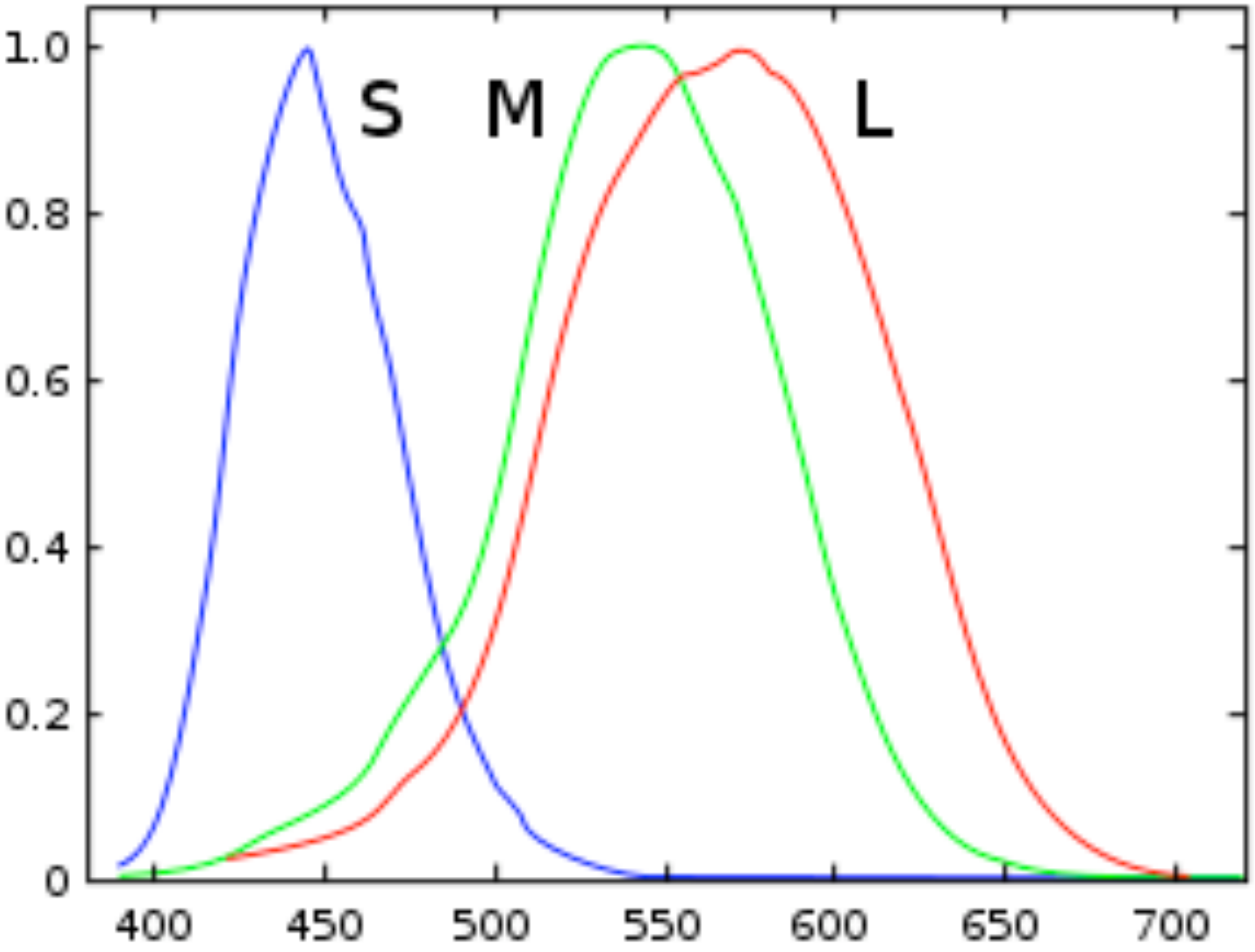
- Primary reason: generate color images (RGB)
- GOES-R will have Red 640nm, and Cyan 470nm
- Current methods use lookup tables to predict green then produce RGB

Problem:

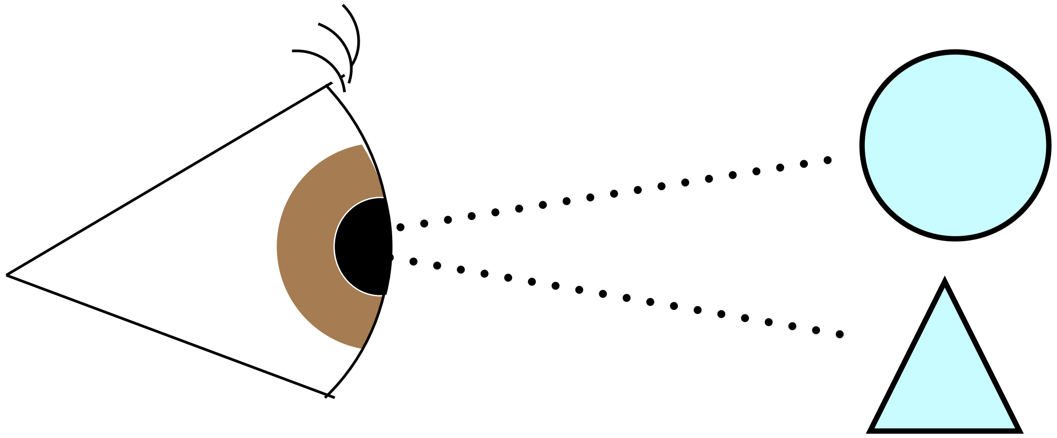
Human color vision **not** based on narrow band RGB

Vision: Wide Band Response





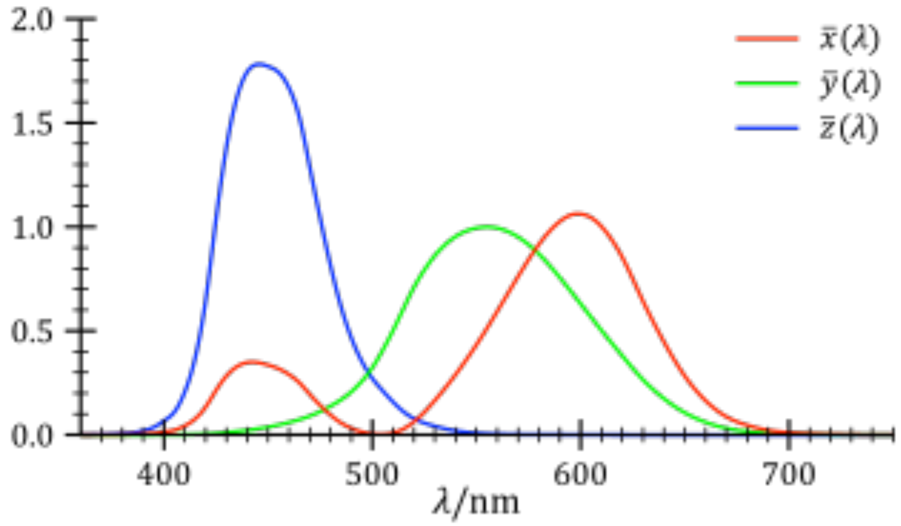
Tristimulus and XYX



I(λ) spectral power distribution

I'(λ)

Two objects have same color <=> XYZ=X'Y'Z'



$$X = \int_0^\infty I(\lambda)\,\overline{x}(\lambda)\,d\lambda$$

$$Y = \int_0^\infty I(\lambda)\,\overline{y}(\lambda)\,d\lambda$$

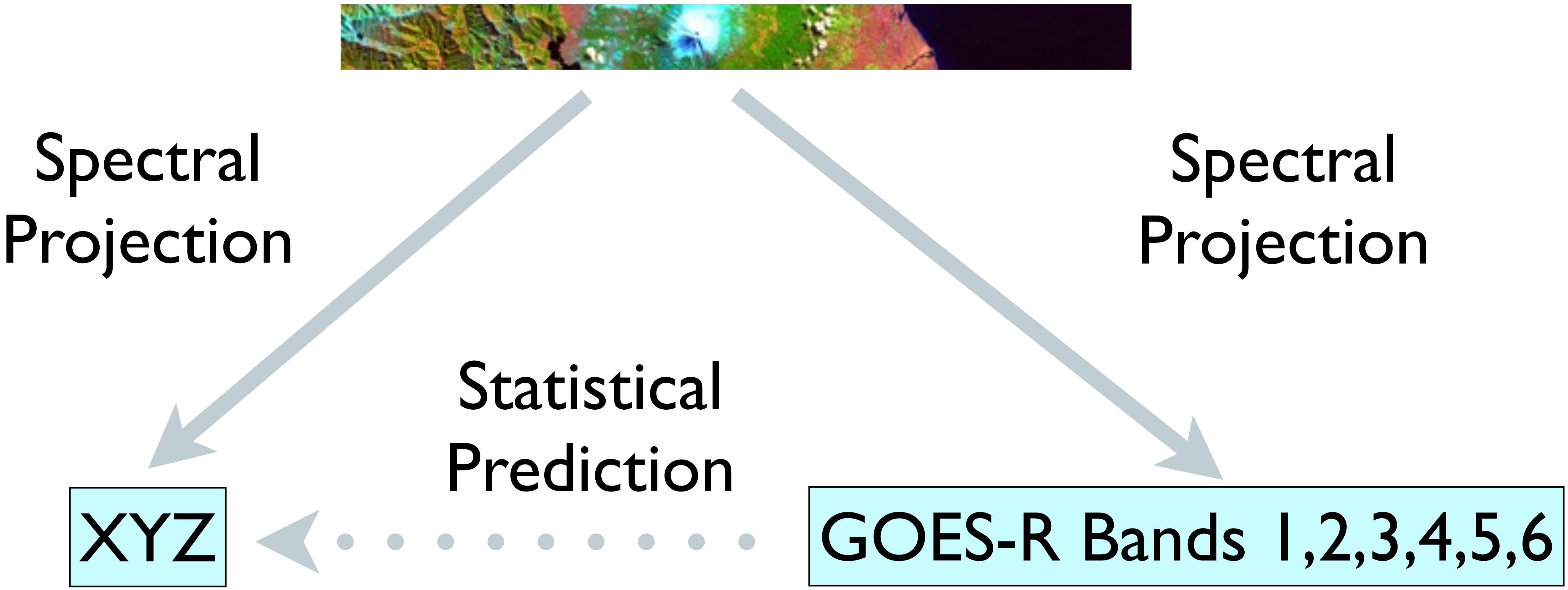
$$Z = \int_0^\infty I(\lambda)\,\overline{z}(\lambda)\,d\lambda$$

Don't estimate green!

Estimate XYZ and get accurate RGB

Hyperion as Spectrometer

Hyperion Data, 220 bands



Spectral Projection

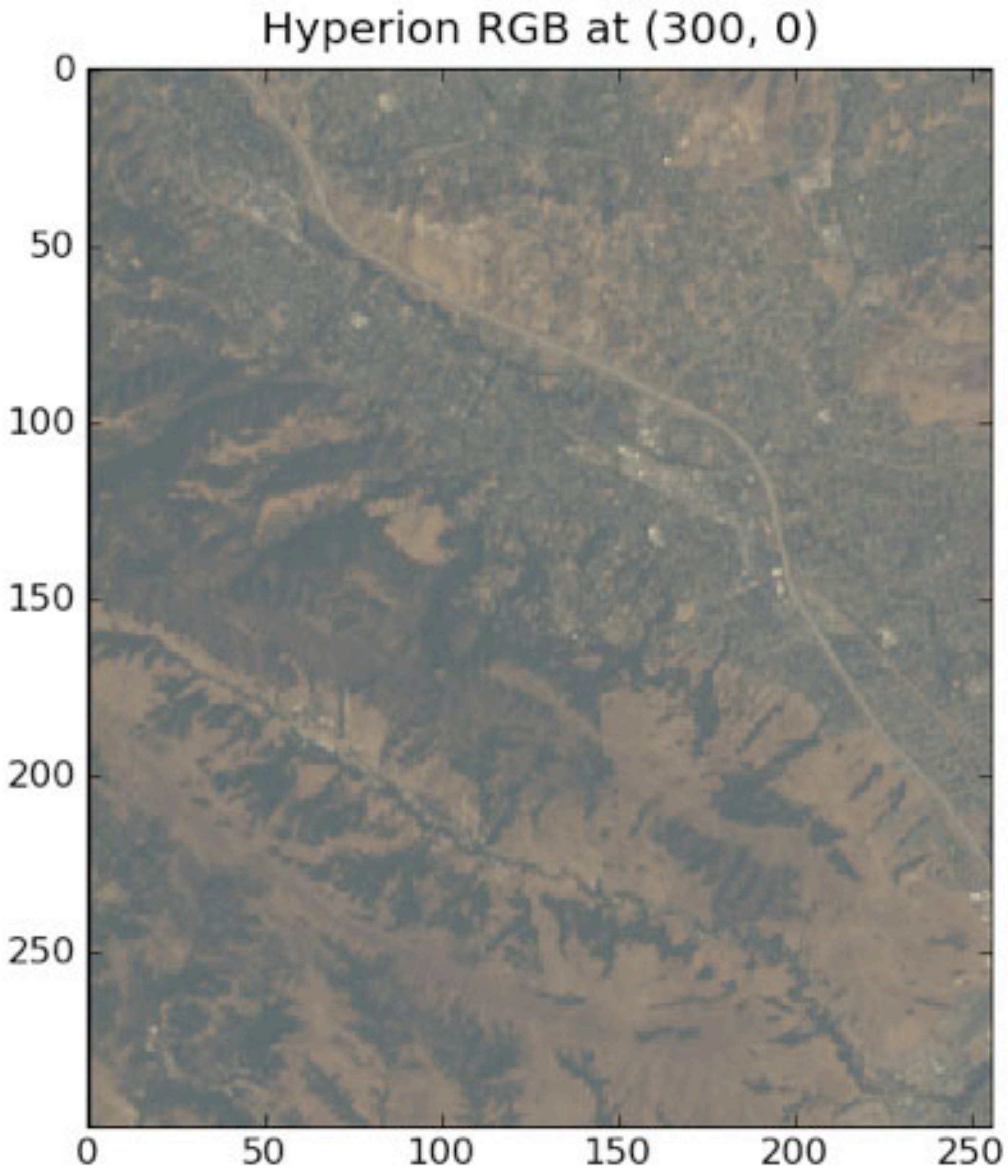
Spectral Projection

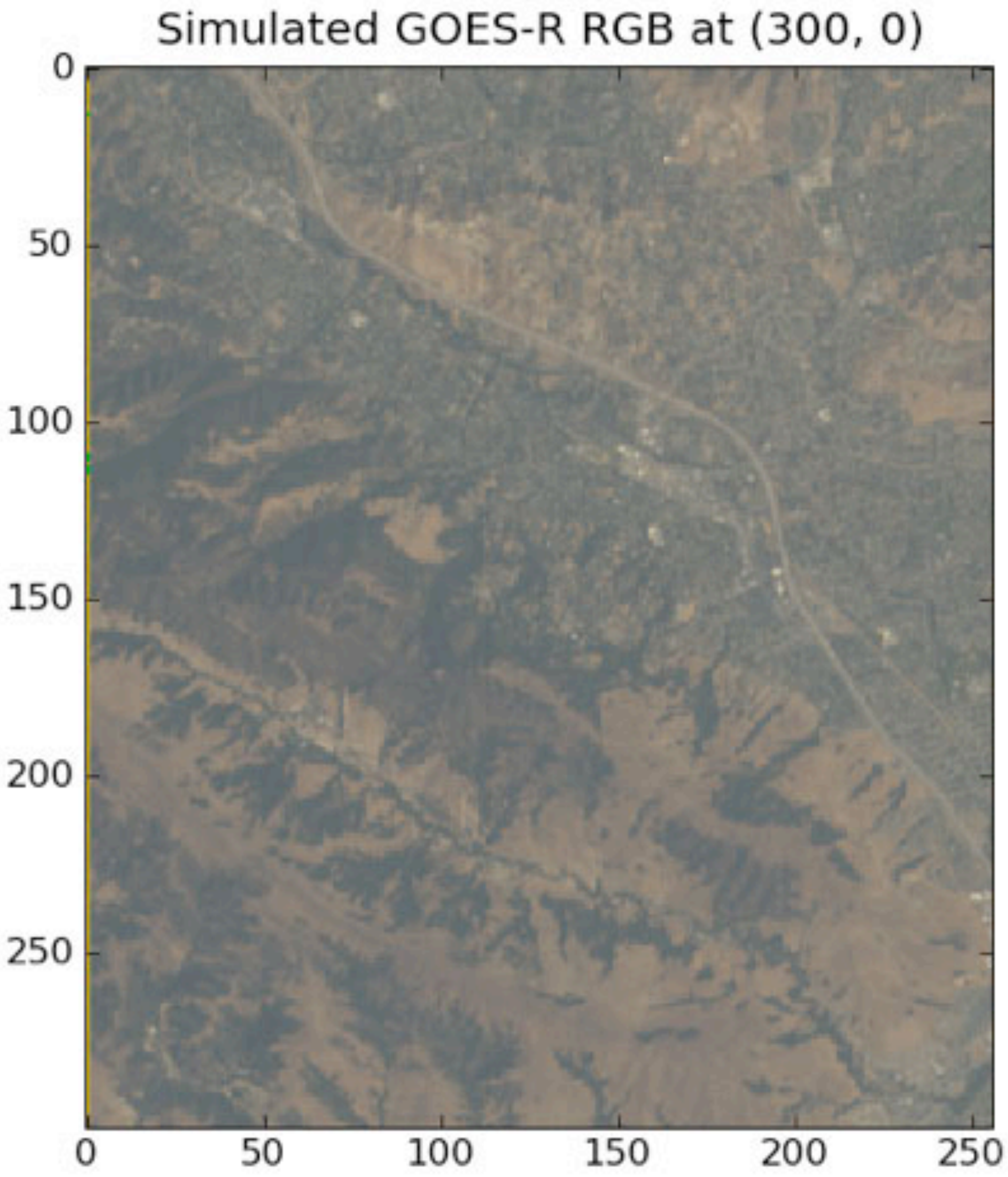
Statistical Prediction

XYZ

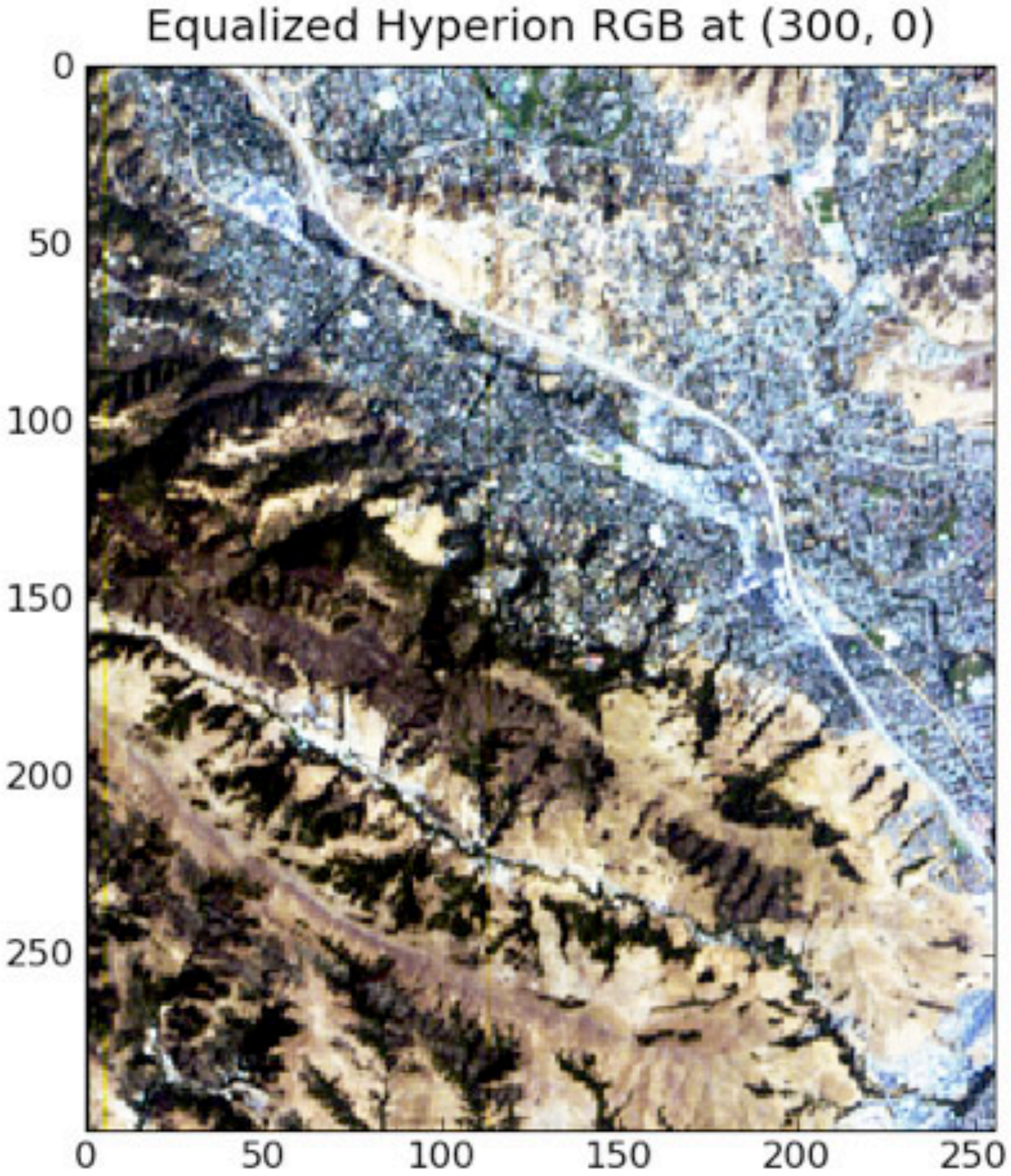
GOES-R Bands 1,2,3,4,5,6

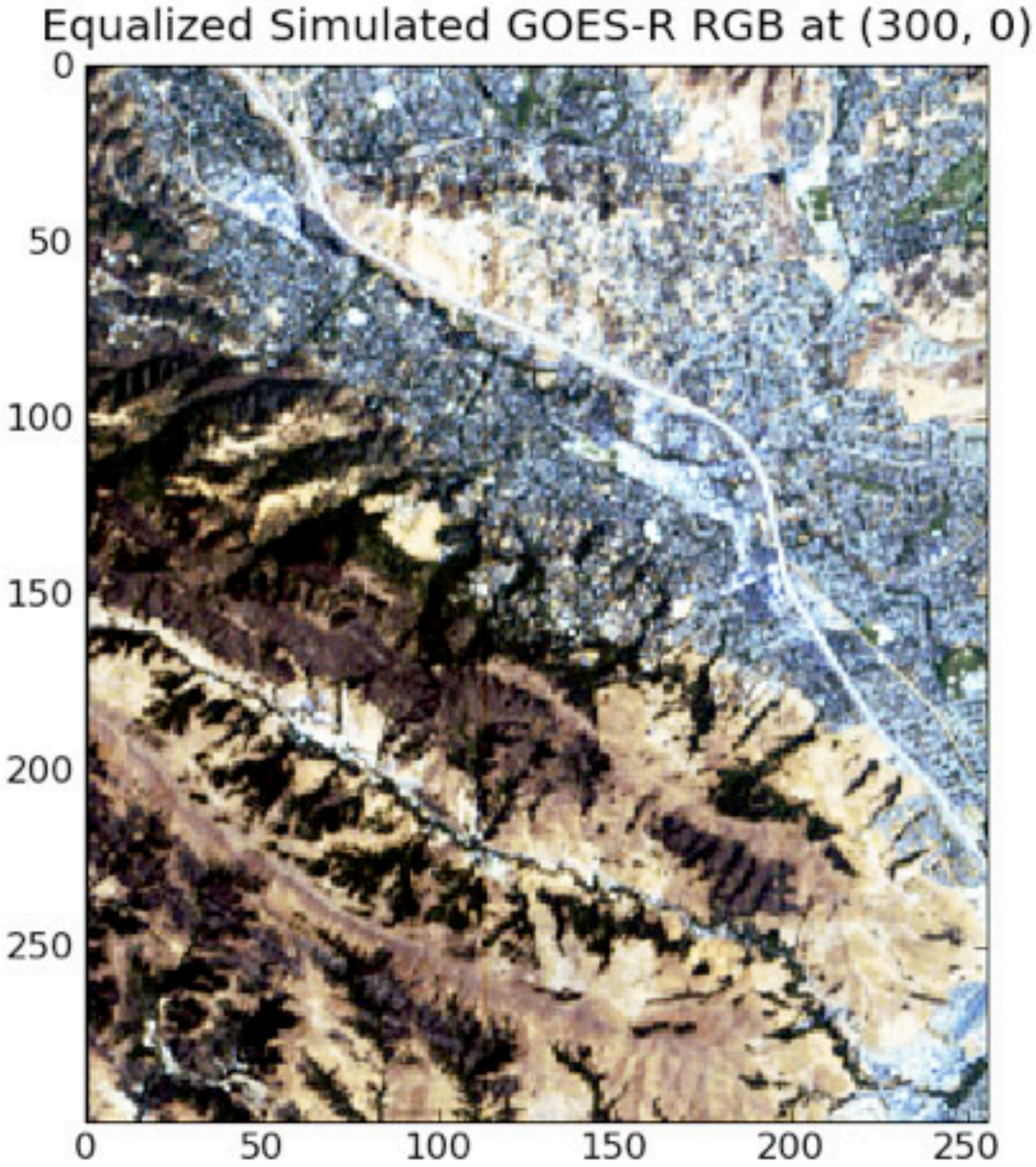
Proof of Concept Results





Equalized Images

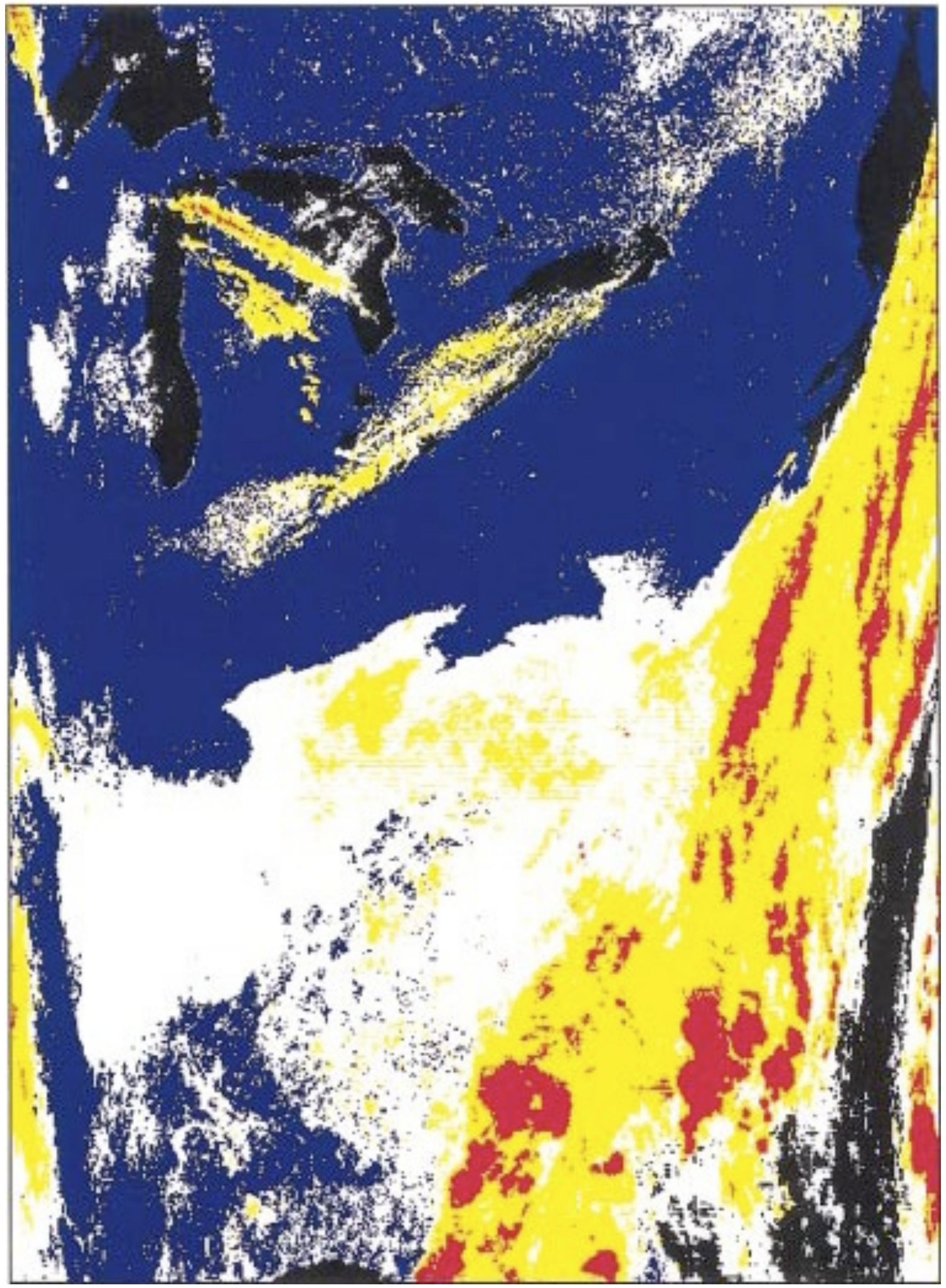




Equalization simply for magnifying differences

Beyond Prediction

- Statistical Estimation applies to clustering and classification tasks
- Example Clustering Problem (from Paul Menzel)
  - What bands are most important for separating different cloud states?
  - How do statistical clusters with those predicted by physics models?



Library of Algorithms

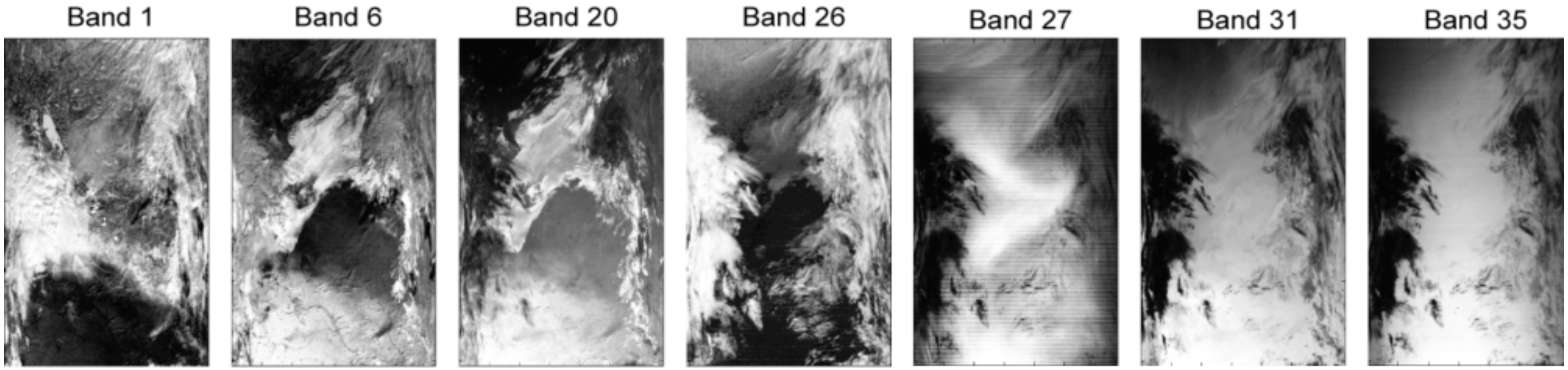
- Many different statistical clustering algorithms
- Hard to evaluate: what defines a good cluster?
- We built a library: implements/wraps major clustering algorithms

Available Clustering Algorithms

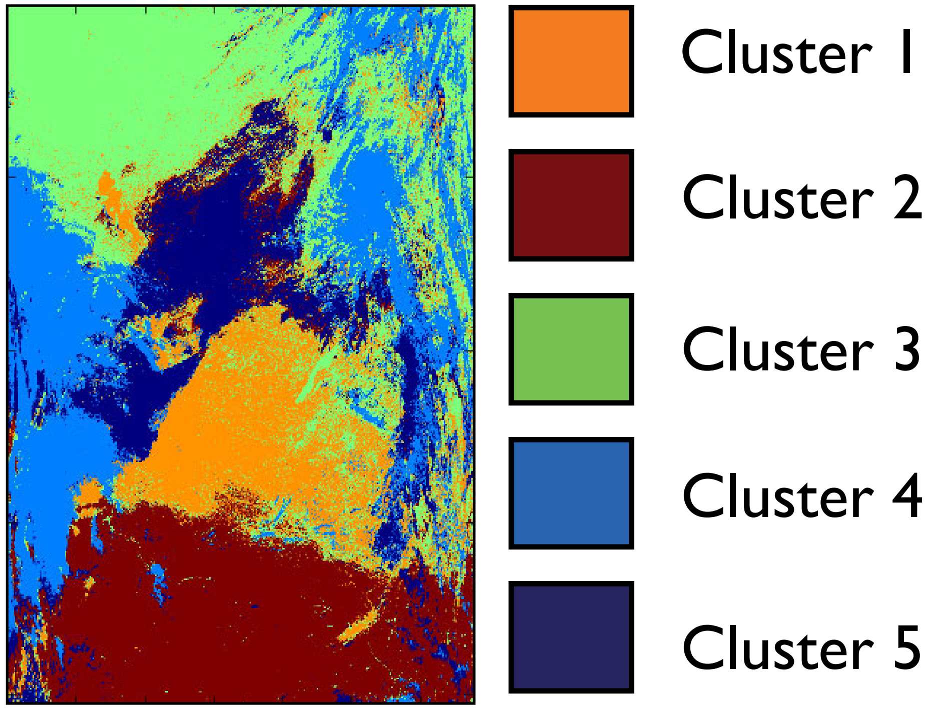
- Agglomerative
- Agglomerative Hierarchical
- Average Link
- Best One Element Move Consensus
- Best of K Consensus
- CC Average Link
- CC Pivot
- Competitive Learning
- Connected Component
- Connected Components
- Expectation Maximization
- Fuzzy K-means
- Graph Cut
- Hierarchal Dimensionality Reduction
- K-means
- Leader Follower
- Majority Rule Consensus
- Mean Shift
- Multi-Dimensional Scaling
- Spectral Clustering
- Stepwise Optimal Hierarchical

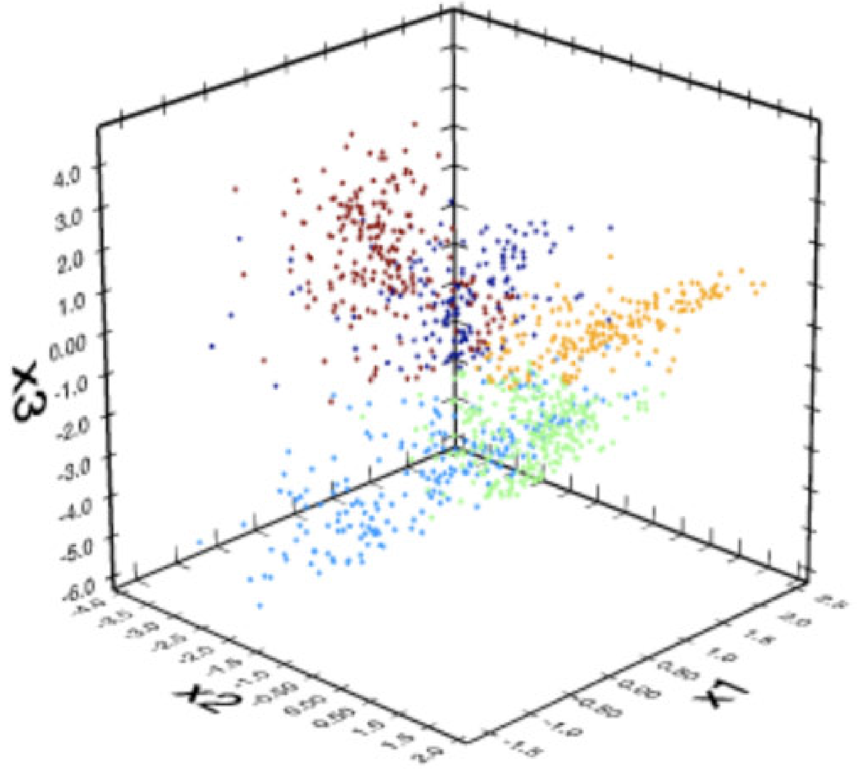
Eg: Competitive Learning

MODIS

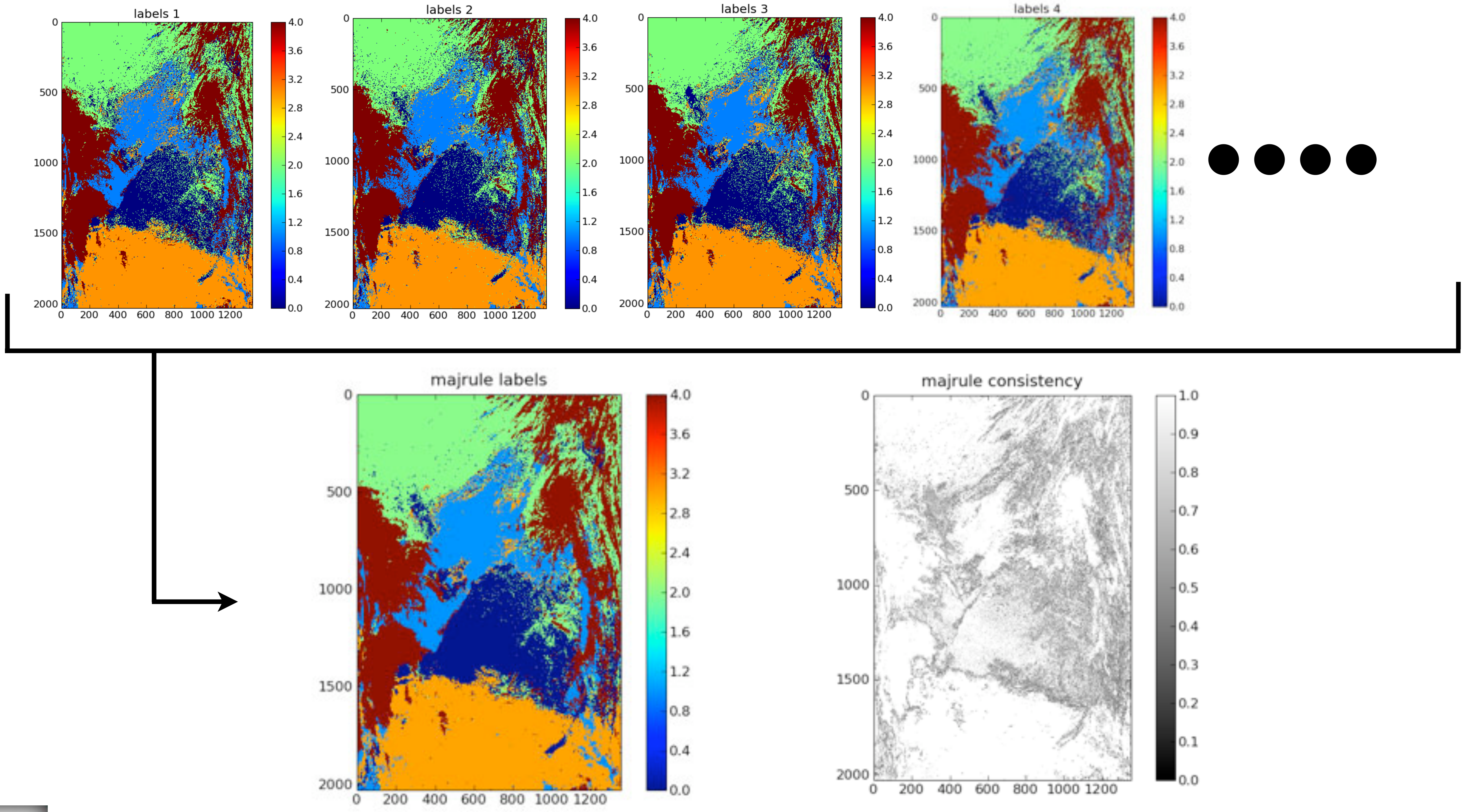


Input: 7 dimensions/pixel





Consensus Clustering



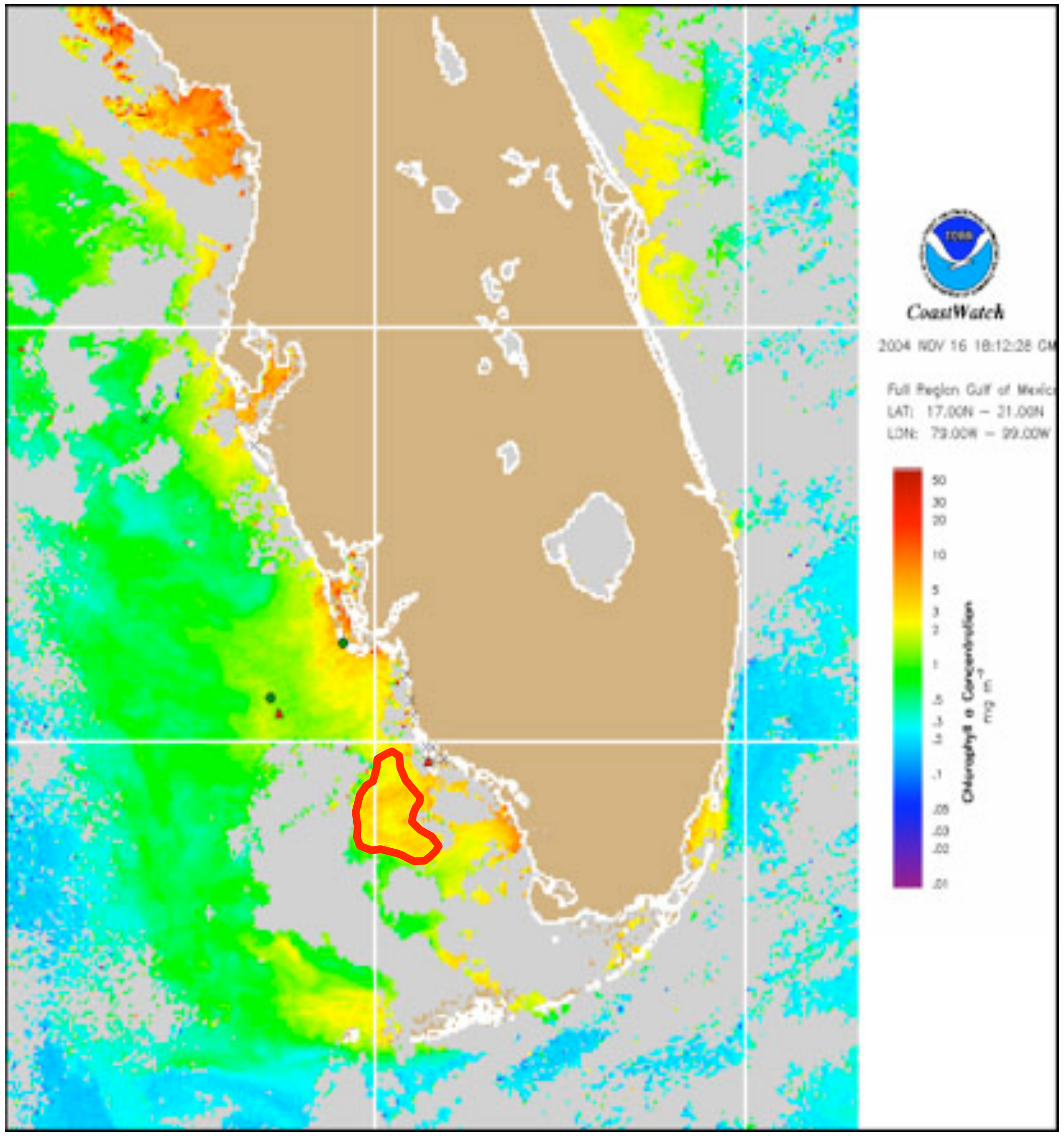
Consensus Label

Agreement

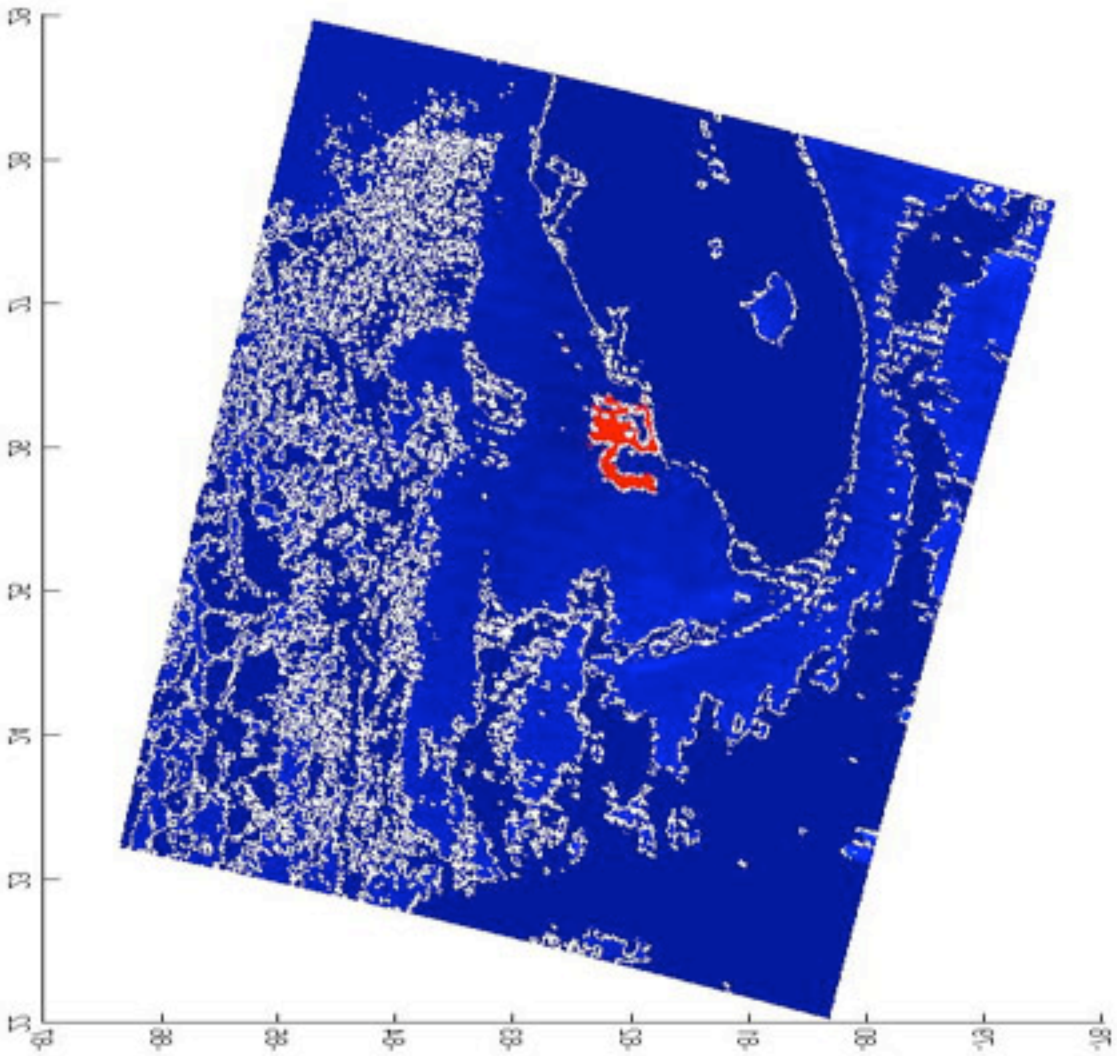
Data Clustering For Classification

Algae Blooms

What multispectral signatures correlate with presence of a bloom?

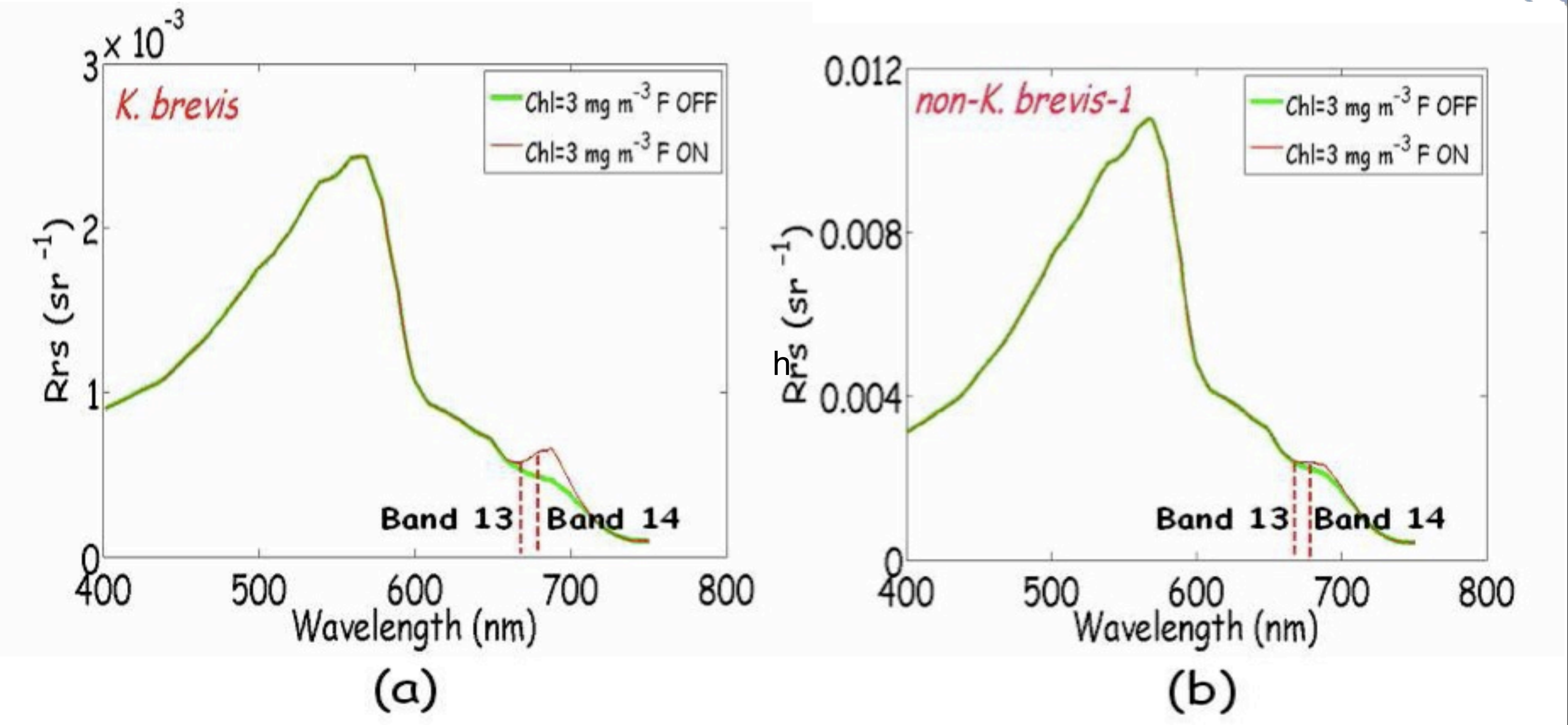


Algae Bloom Bulletin: bloom outlined in red



Clustering Result: bloom shown in red

Modeled Remote Sensing Reflectance Spectra

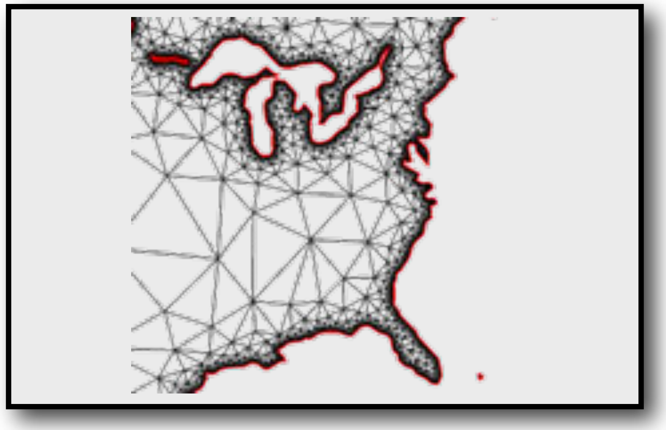


The solid green spectra are when chlorophyll fluorescence is excluded from the simulation and solid red spectra are when fluorescence is included in the simulation assuming 0.75% quantum yield. Band 13 and 14 are MODIS bands centered at 667nm and 678nm respectively.

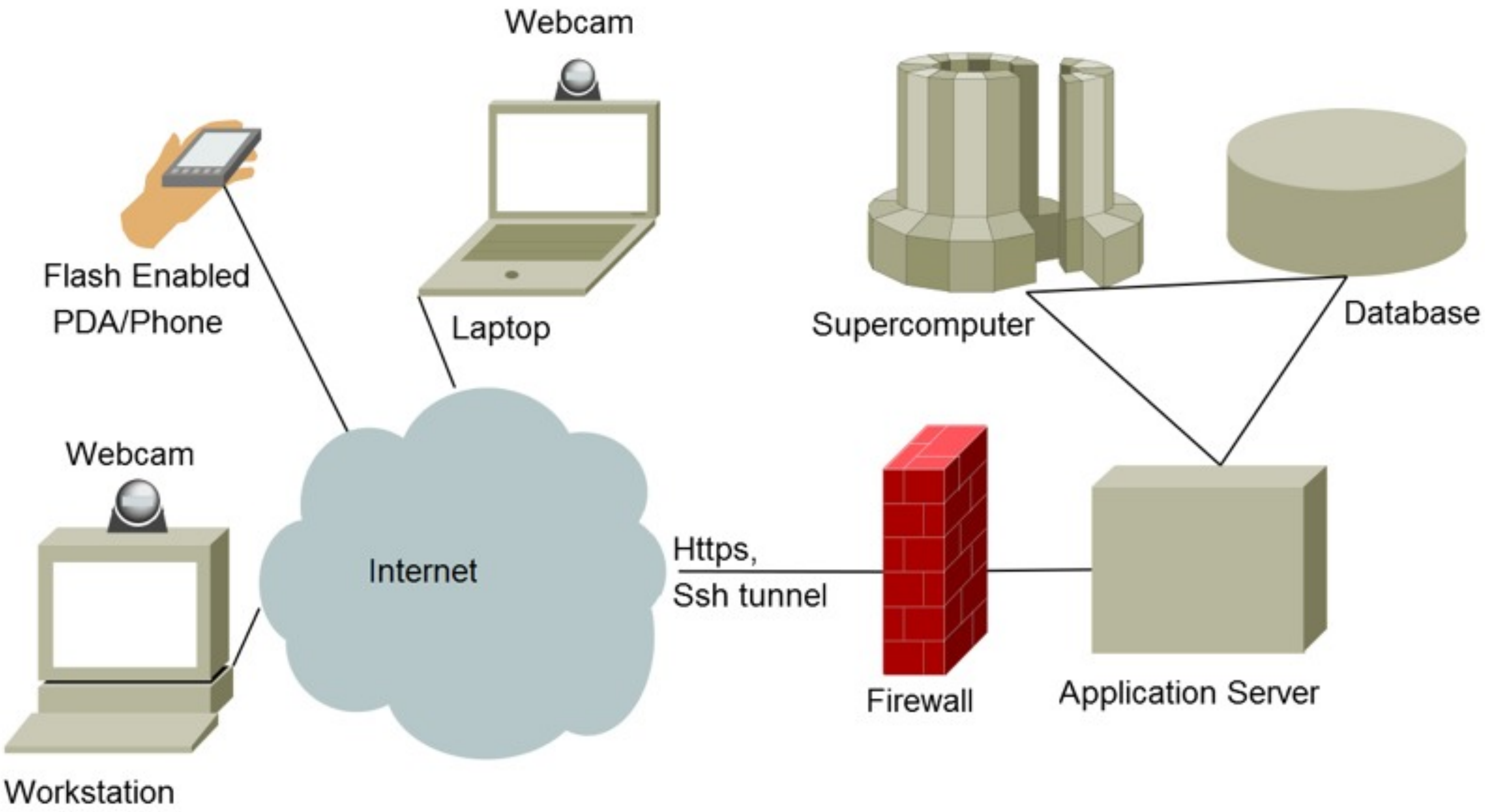
S. Ahmet et. al. “Novel optical techniques for detecting and classifying toxic dinoflagellate *Karenia brevis* blooms using satellite imagery”

Graphyte Tool Kit

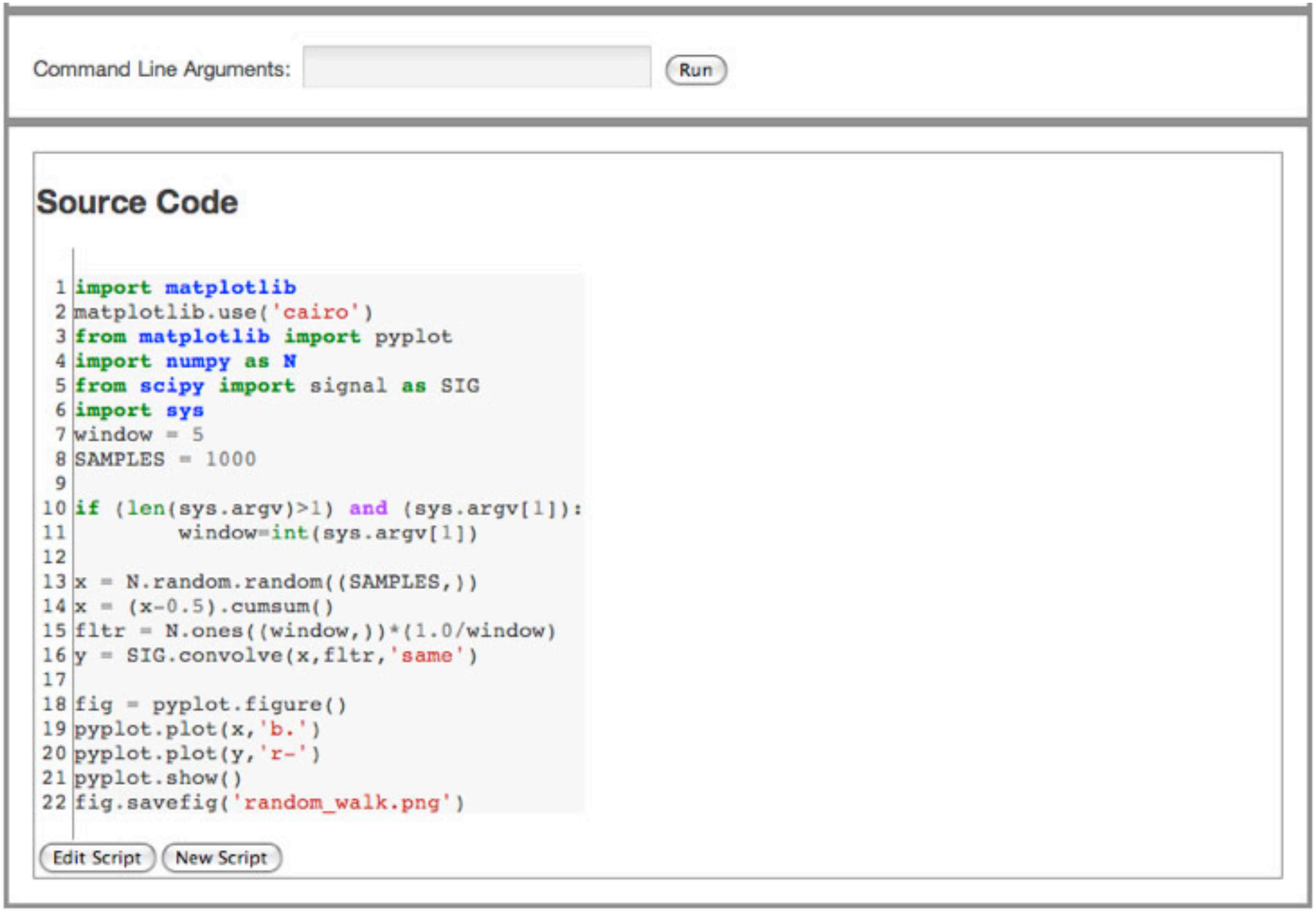
- Web based interface to:
  - Data
  - Computation
  - Algorithms
- 2D/3D graphical interactive tools
  - Data Exploration
  - Data Visualization



Hardware Architecture



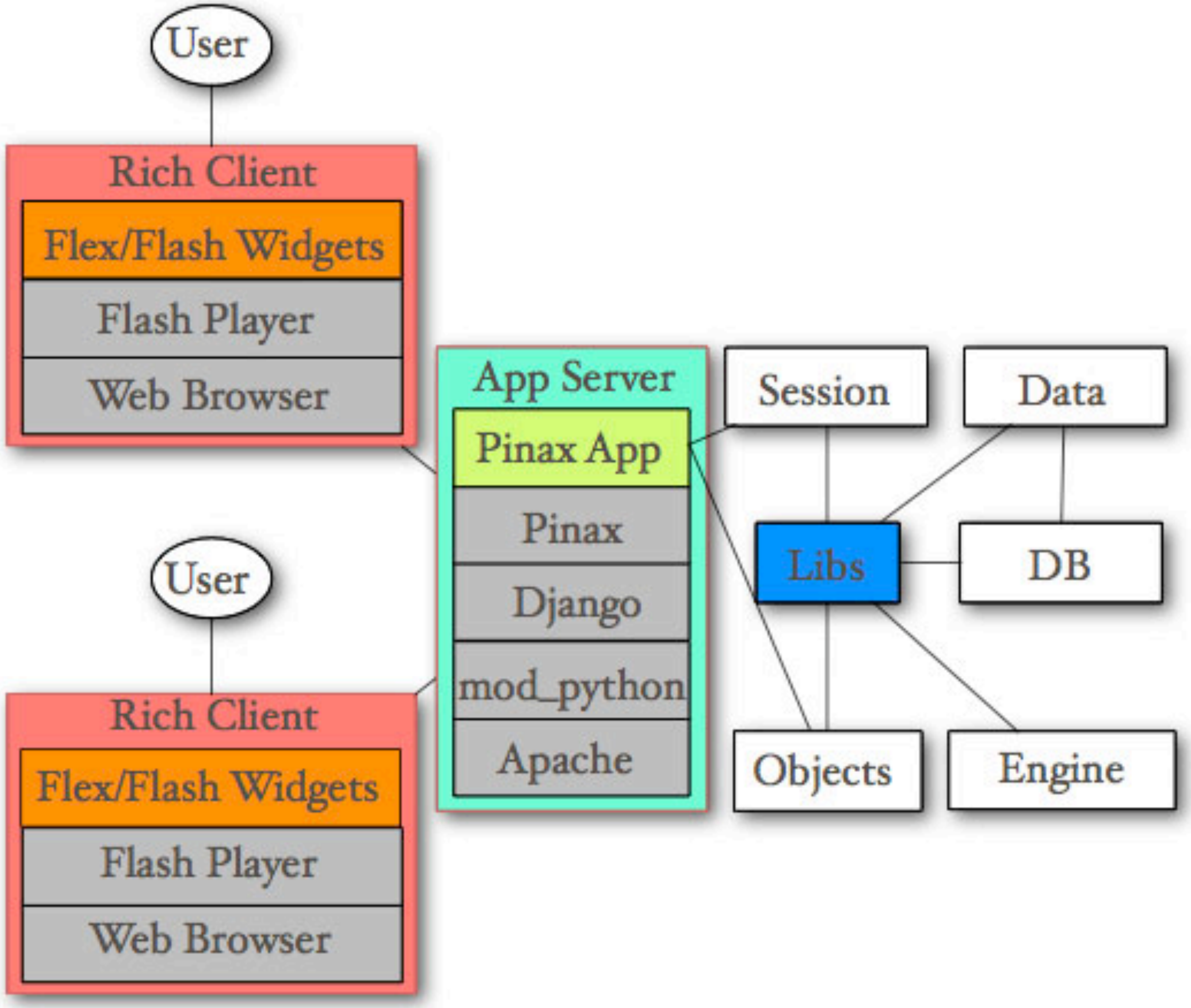
Edit Code In Browser



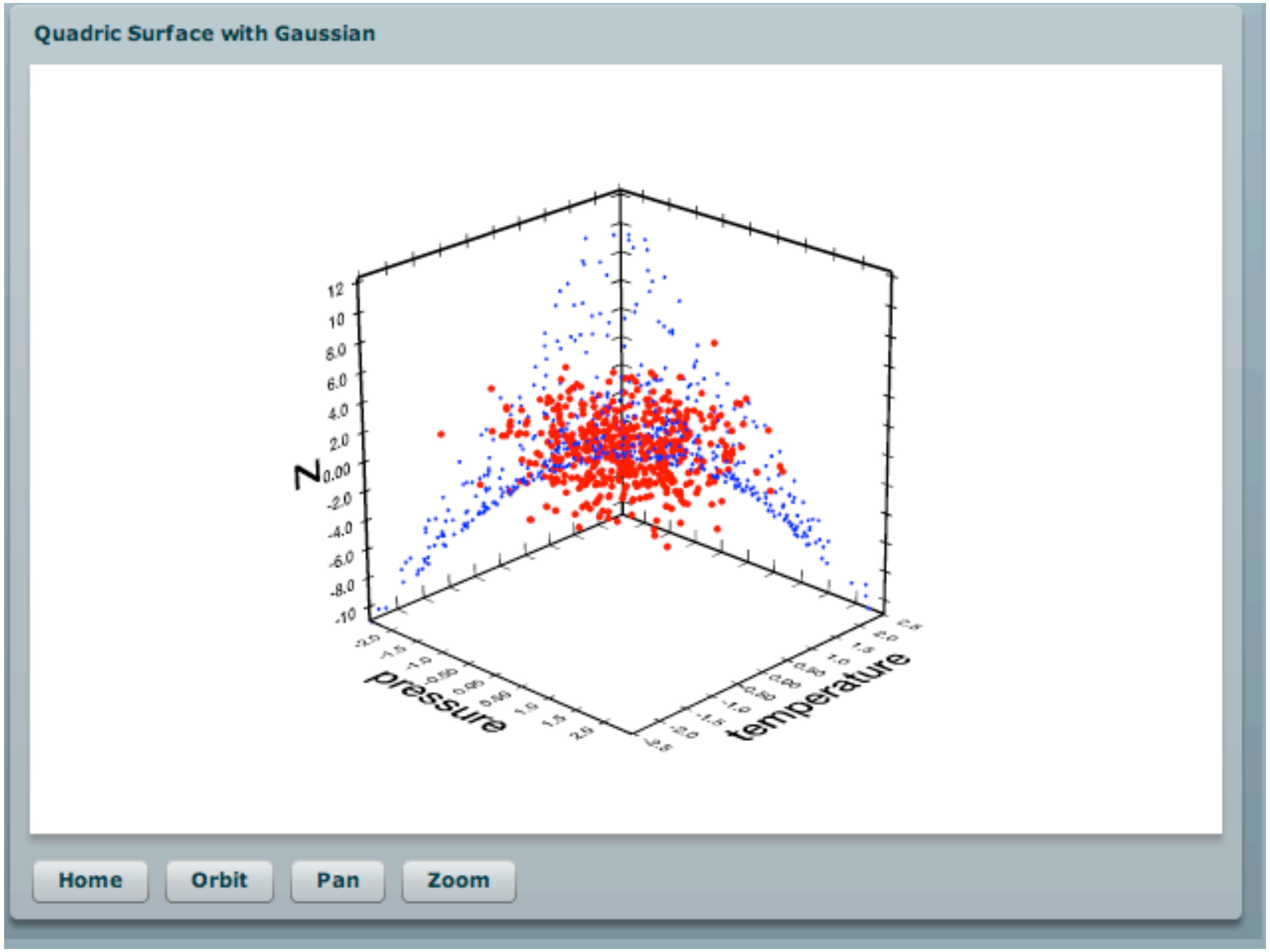
Source Code

```
import matplotlib
matplotlib.use('cairo')
from matplotlib import pyplot
import numpy as N
from scipy import signal as SIG
import sys
window = 5
SAMPLES = 1000

if (len(sys.argv)>1) and (sys.argv[1]):
        window=int(sys.argv[1])

x = N.random.random((SAMPLES,))
x = (x-0.5).cumsum()
fltr = N.ones((window,))*(1.0/window)
y = SIG.convolve(x,fltr,'same')

fig = pyplot.figure()
pyplot.plot(x,'b.')
pyplot.plot(y,'r-')
pyplot.show()
fig.savefig('random_walk.png')
```

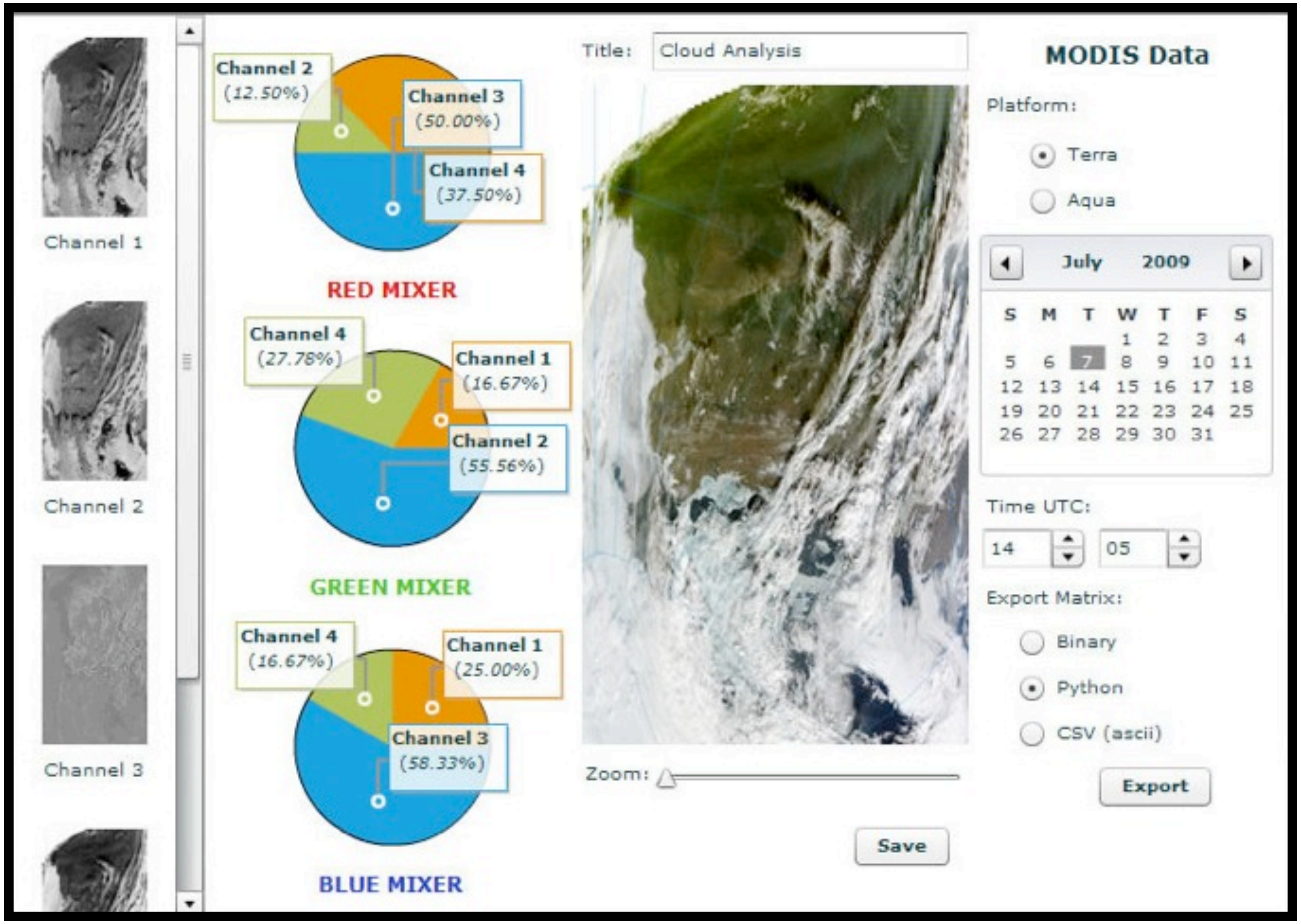
Software Architecture



Interactive 3D Scatter Plot



Rich Internet Application



Title: Cloud Analysis

Platform:

● Terra

○ Aqua

July 2009

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

Time UTC: 14 05

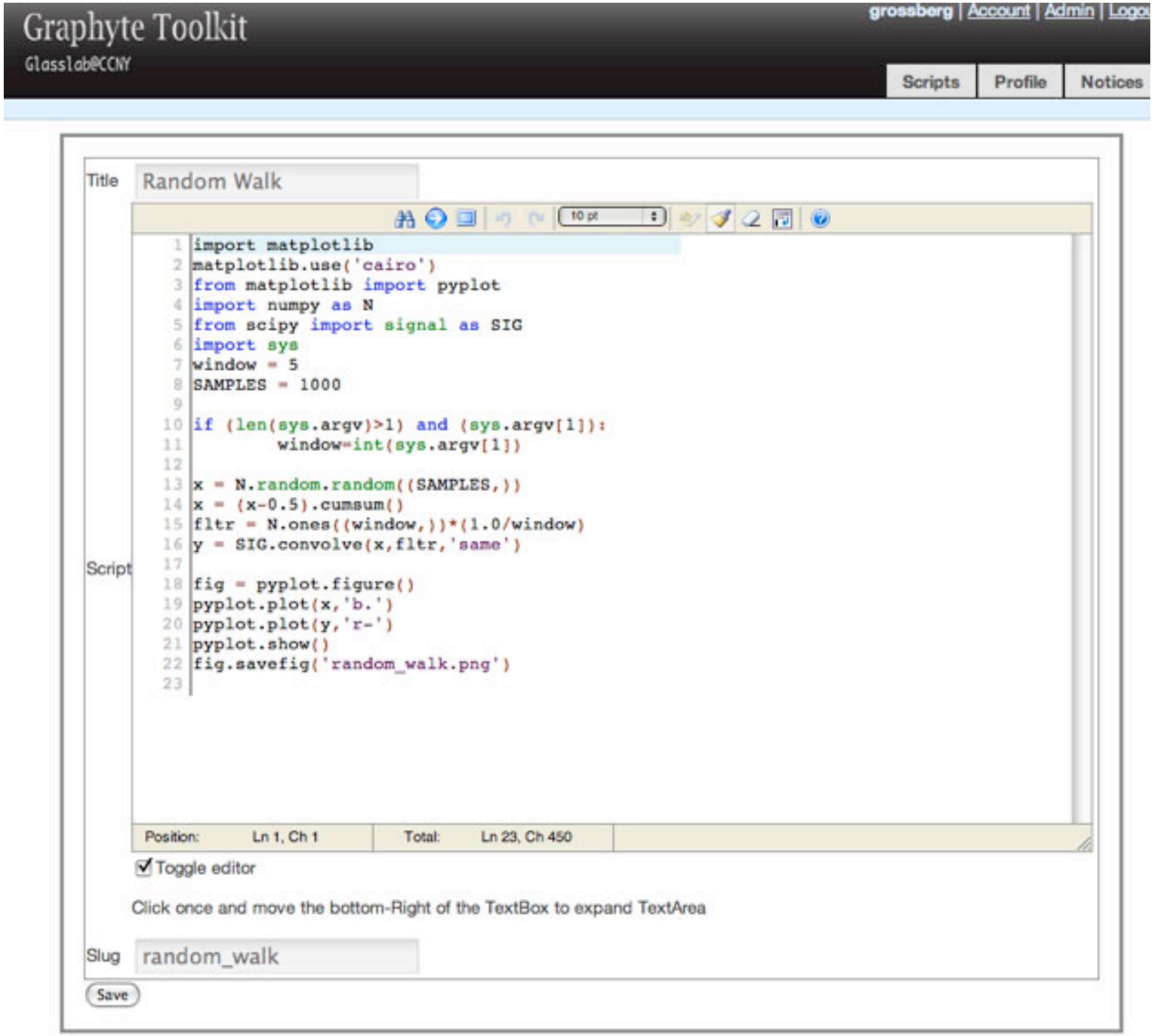
Export Matrix:

○ Binary

● Python

○ CSV (ascii)

Edit/Run Code Through Browser



Title: Random Walk

Script

```python
import matplotlib
matplotlib.use('cairo')
from matplotlib import pyplot
import numpy as N
from scipy import signal as SIG
import sys
window = 5
SAMPLES = 1000

if (len(sys.argv)>1) and (sys.argv[1]):
        window=int(sys.argv[1])

x = N.random.random((SAMPLES,))
x = (x-0.5).cumsum()
fltr = N.ones((window,))*(1.0/window)
y = SIG.convolve(x,fltr,'same')

fig = pyplot.figure()
pyplot.plot(x,'b.')
pyplot.plot(y,'r-')
pyplot.show()
fig.savefig('random_walk.png')
```

Position: Ln 1, Ch 1 Total: Ln 23, Ch 450

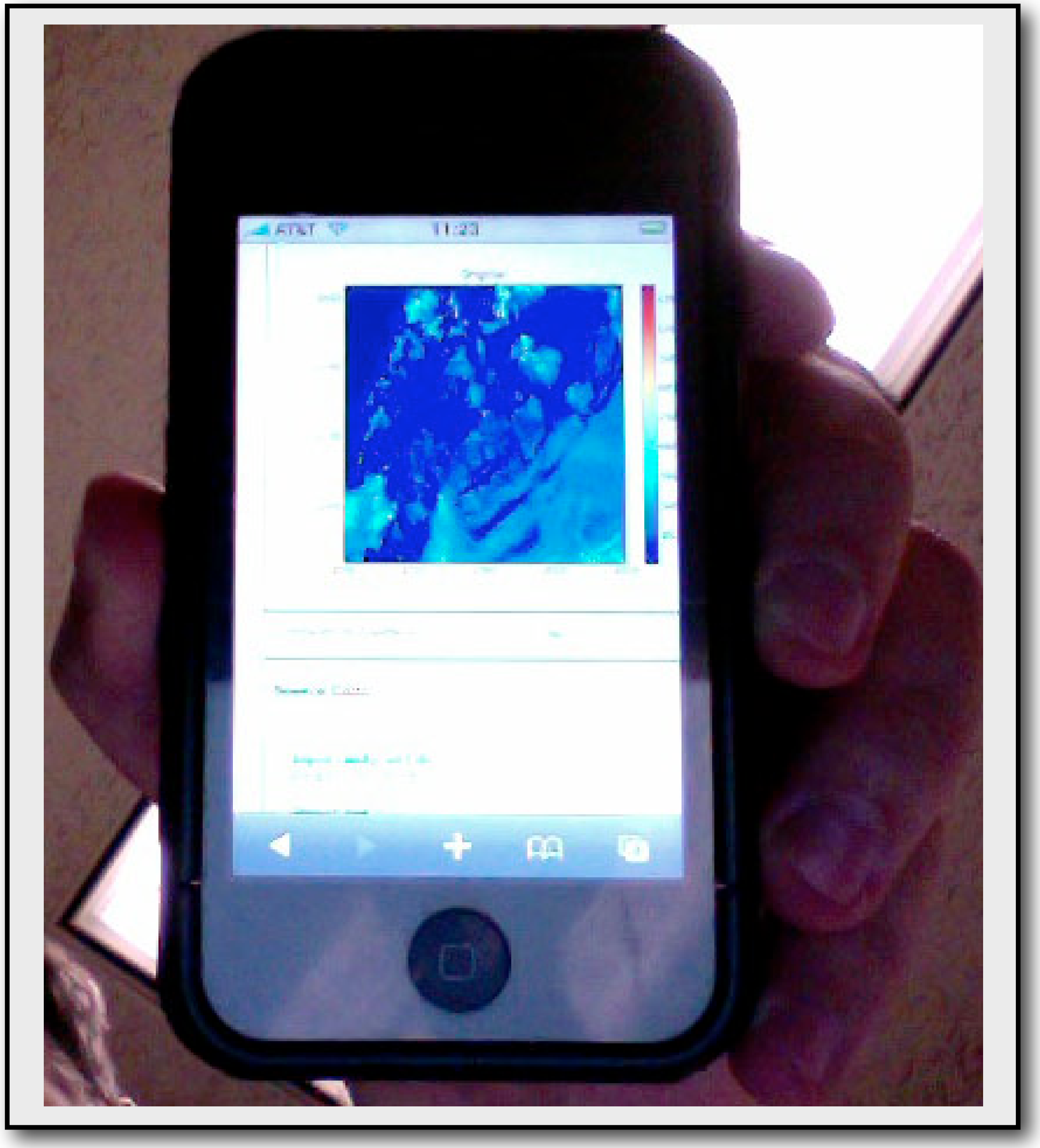
☑ Toggle editor

Click once and move the bottom-Right of the TextBox to expand TextArea

Slug: random_walk

Save

Near Universal Availability



Conclusion

- Provide Expertise
  - High Dimensionality
  - Large Data Sets
  - Statistical Clustering, Estimation, Classification
- Provide Tools for
  - Computation
  - Data Access
  - Visualization
  - Remote Collaboration