
National Aeronautics and Space Adr
NASA
NASA Ames Data Sciences Group
Nikunj C. Oza, Ph.D.
Leader, Data Sciences Group
nikunj.c.oza@nasa.gov
www.nasa.gov
•1

The Data Sciences Group at NASA Ames

Data Mining Research and Development (R&D) for application to NASA problems (Aeronautics, Earth Science, Space Exploration, Space Science)

Group Members

Ilya Avrekh
Kamalika Das, Ph.D.
Dave Iverson
Vijay Janakiraman, Ph.D.
Rodney Martin, Ph.D.
Bryan Matthews
David Nielsen
Nikunj Oza, Ph.D.
Veronica Phillips
John Stutz
Hamed Valizadegan, Ph.D.
+ summer students

Funding Sources

- Science Mission Directorate: AIST and CMAC programs
- NASA Aeronautics Research Mission Directorate- ATD, SMART-NAS, SASO Project
- NASA Engineering and Safety Center
- Exploration Systems Mission Directorate, Exploration Technology Development Program
- Non-NASA: DARPA, DoD

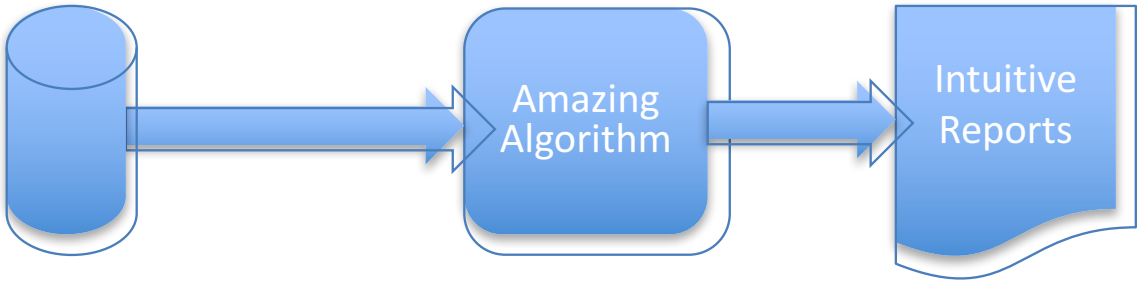
Team Members are NASA Employees, Contractors, and Students.

Example Data Mining Problems

- Aeronautics: Anomaly Detection, Precursor Identification, text mining (classification, topic identification)
- Earth Science: Filling in missing measurements, anomaly detection, teleconnections, climate understanding
- Space Science: Kepler planet candidates
- Space Exploration: system health management, vascular structure identification

Four V's of ~~Big~~ Tough, Sleep Depriving Data

- Volume:
  - Radar Tracks: 47 facilities (1 year) ~423 GB (Compressed), ~3.2 TB (CSV)
  - Weather and Forecast (Entire NAS): CIWS ~2.8 TB
- Velocity
  - Radar Tracks: 47 Facilities
    - ~35 GB/month (compressed).
    - ~268 GB/month (uncompressed)
  - Weather and Forecast (Entire NAS): CIWS ~233 GB/month
- Veracity
  - Data drop outs
  - Duplicate tracks
  - Track ending in mid air
  - Reused flight identifiers
- Variety
  - Numerical (continuous/binary)
  - Weather (forecast/actual)
  - Radar/Airport meta data
  - ATC Voice
  - ASRS text reports (Pilot/Controller)



Aeronautics Data Mining Problems

- Anomaly Detection
  - Anomaly Discovery over large set of variables
  - Particular variable of interest, for example, fuel burn
    - Determine expected instantaneous fuel burn given current state of aircraft
    - Compare with actual instantaneous fuel burn
    - Where difference is high, problem may be occurring
- Precursor Identification
  - Given undesirable effect (e.g., go-around), identify precursors (e.g., overtake situation, high speed approach)
- Text mining
  - Text classification, topic identification

Topic Extraction Example

TOPIC 1	TOPIC 2	TOPIC 3
autoplt	time	apch
acft	day	rwy
spd	leg	visual
capture	contributing	ils
mode	factors	twr
rate	hrs	lndg
level	crew	loc
engaged	factor	arpt
leveloff	fatigue	final
vert	night	missed
ctl	trip	clred
disconnected	rest	msl
selected	duty	intercept
fpm	flying	vectored
light	long	sight
clb	late	gar
pitch	previous	terrain
manually	incident	field
warning	lack	uneventful
pwr	alerter	ctl

Other examples of 'fatigue'

Altitude Deviation
Spatial Deviation
Ramp Excursion
Landing without clearance
Runway Incursion
Unstable Approach

Aeronautics Anomaly Detection: Current Methods

Exceedance-Based Methods

- Known anomalies
- Conditions over 2-3 variables (e.g., speed > 250 knots, altitude = 1000 ft, landing)
- Cannot identify unknown anomalies
- Low false positive rate, high false negative (missed detection) rate.

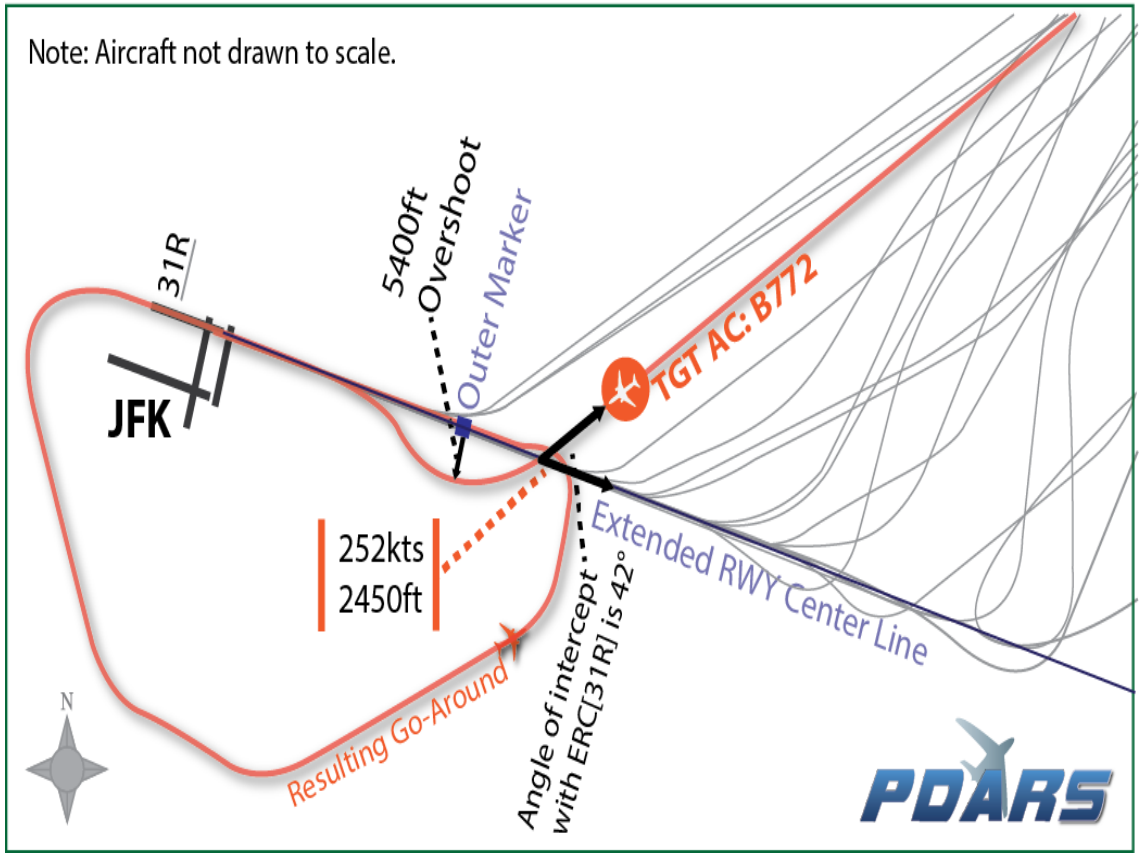
Data-Driven Methods

- DISCOVER anomalies by
  - learning statistical properties of the data
  - finding which data points do not fit (e.g., far away, low probability)
  - no background knowledge on anomalies needed
- Complementary to existing methods
  - Low false negative (missed detection) rate
  - Higher false positive rate (identified points/flights unusual, but not always operationally significant)
- Data-driven methods -> insights -> modification of exceedance detection

Example: High Speed Go-Around

- Overshoots Extended Runway Centerline (ERC) by over 1 SM

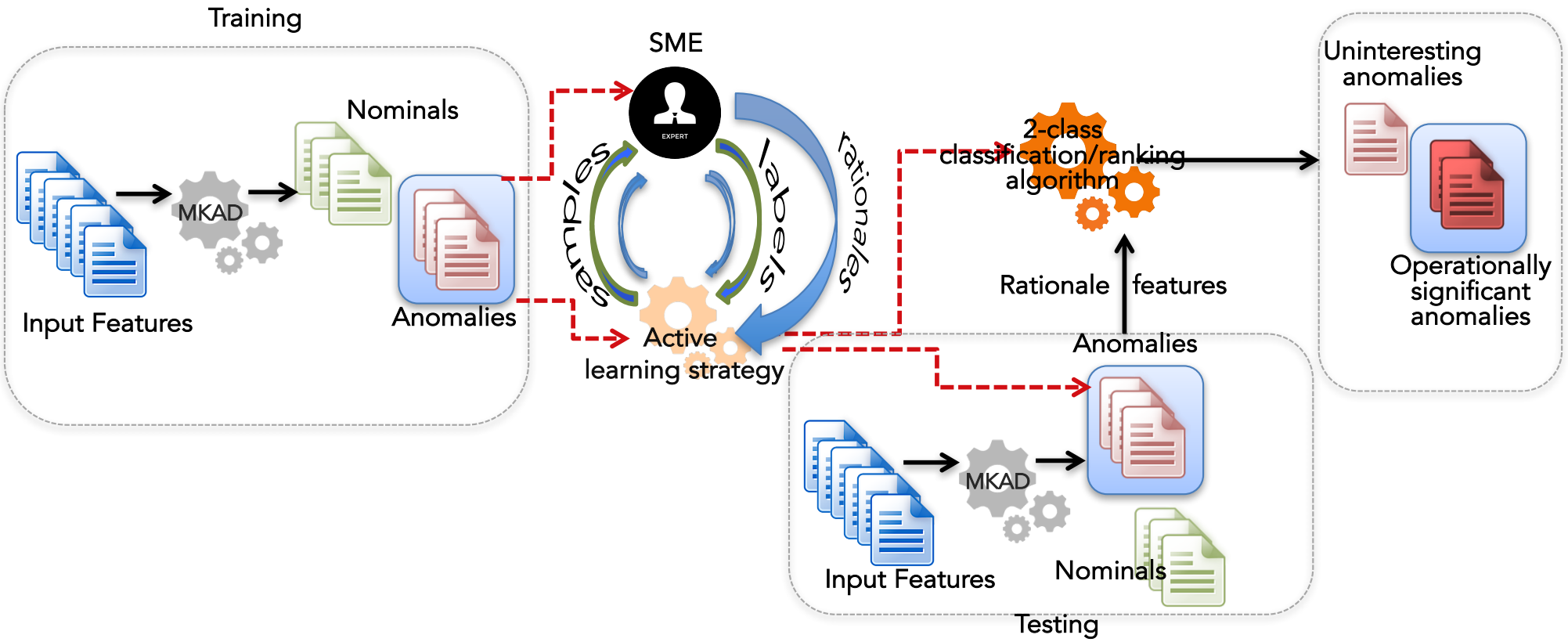
- Over 250 Kts @2500 Ft.
- Angle of intercept > 40°
- Overshoots 2nd approach



Bryan Matthews, David Nielsen, John Schade, Kennis Chan, and Mike Kiniry, Automated Discovery of Flight Track Anomalies, *33rd Digital Avionics Systems Conference*, 2014

Providing Domain Expert Feedback

Active learning with rationales framework



Manali Sharma, Kamalika Das, Mustafa Bilgic, Bryan Matthews, David Nielsen, and Nikunj Oza, Active Learning with Rationales for Identifying Operationally Significant Anomalies in Aviation, *European Conference on Machine Learning and Principles and Practices Of Knowledge Discovery (ECML-PKDD)*, 2016

Earth Science Example

- Understand relationships between ecosystem dynamics and climatic factors
- Model as a regression analysis problem
- 3 science questions
  - Magnitude and extent of ecosystem exposure, sensitivity and resilience to the 2005 and 2010 Amazon droughts
  - Understand human-induced and other attribution as causes of vegetation anomalies
  - How learned dependency model varies across eco-climatic zones and geographical regions on a global scale

NASA ESTO AIST-14 project, Uncovering Effects of Climate Variables on Global Vegetation (PI: Kamalika Das, Ph.D.)

Problem Formulation

- Point-to-point regression analysis (Genetic Programming based Symbolic Regression)
- Estimate spatio-temporal dependency of forest ecosystems on climate variables

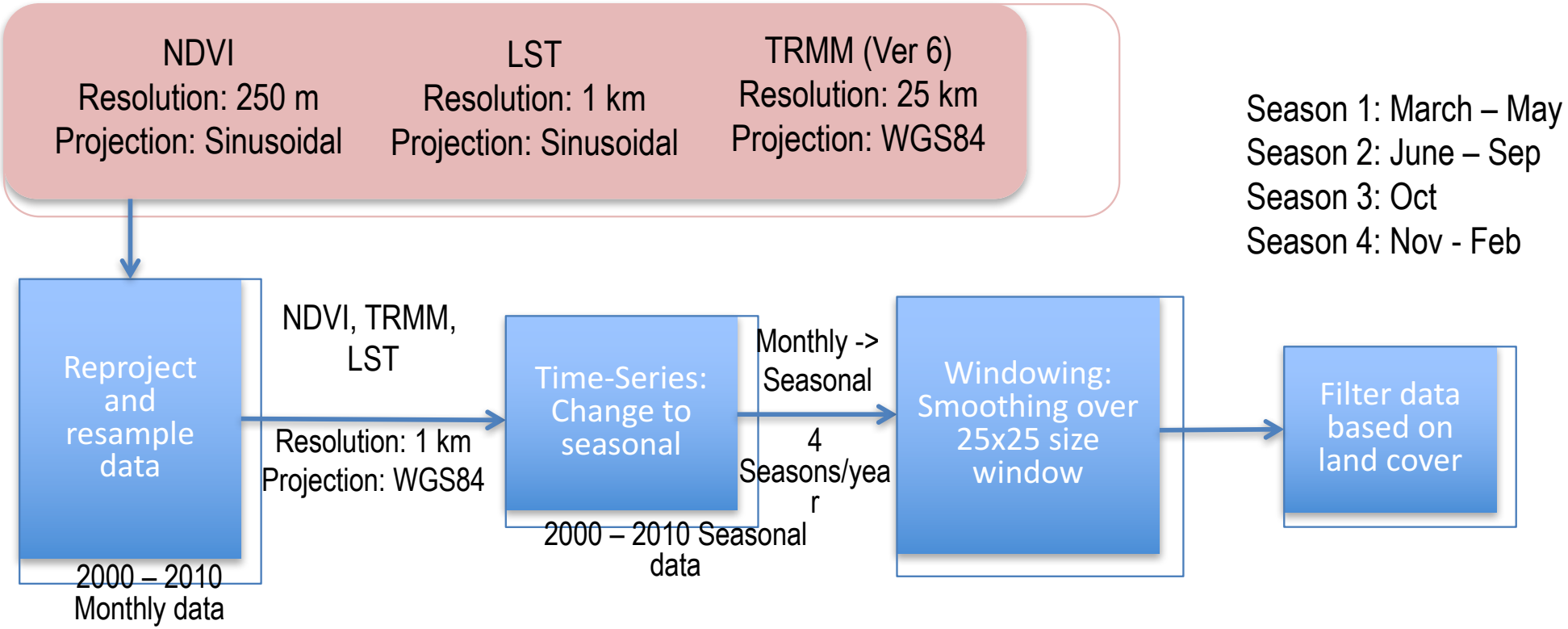
$$V_{ij}^{t} = f(Lc_{ij}, CV_{ij}^{t}, CV_{nb}^{t}, CV_{ij}^{t-1}, CV_{nb}^{t-1}, \ldots\ldots CV_{ij}^{t-k}, CV_{nb}^{t-k})$$

V:vegetation,
LC:landcover type,
CV:climate variable(s)

i,j: pixel location indices
t: time index
nb: spatial neighborhood of index i,j
k: temporal dependency

Open challenges: 1. Estimating function *f*
2. Estimating best choices for *k, nb*

Data Pipeline



Results for 2004-2010

Mean Squared Error

Year	Ridge Regression	LASSO	SVR	Symbolic Regression
2004	0.284	0.284	0.280	0.262
2005	0.289	0.289	0.288	0.278
2006	0.426	0.426	0.430	0.321
2007	0.374	0.374	0.370	0.318
2008	0.308	0.308	0.310	0.336
2009	0.353	0.353	0.360	0.328
2010	0.546	0.547	0.540	0.479

Marcin Szubert, Anuradha Kodali, Sangram Ganguly, Kamalika Das, and Josh C. Bongard, Reducing Antagonism between Behavioral Diversity and Fitness in Semantic Genetic Programming, Proceedings of the Genetic and Evolutionary Computation Conference (GECCO), pp. 797-804, 2016.

Ongoing and Future Work

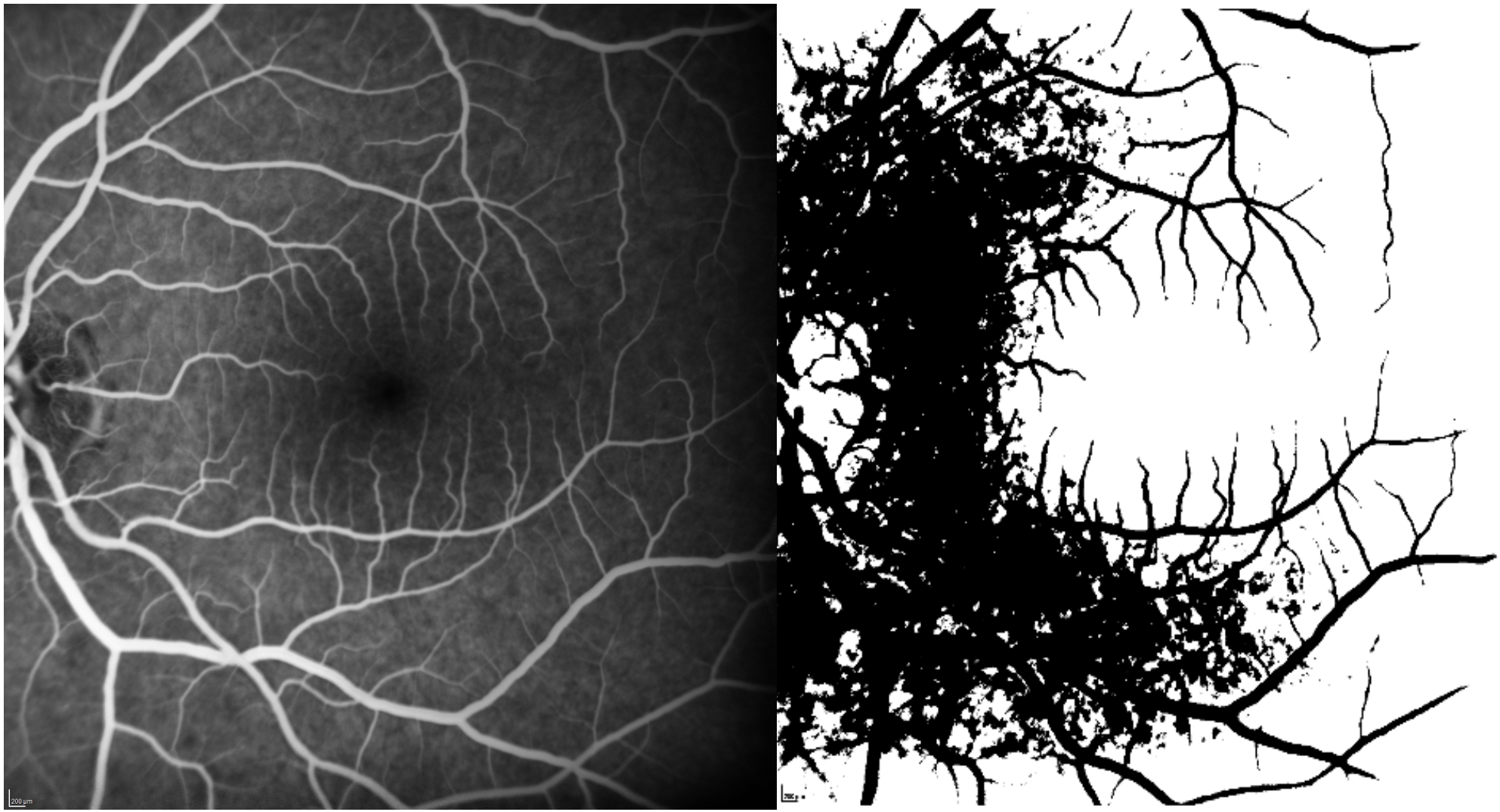
- Experiment with different combinations of temporal lookback and/or spatial effects
- Introduce additional regressors (radiation, forest fire maps, deforestation maps)
- Study the effect of different regressors on different Amazon tiles
- Derive nonlinear GP models on Amazon tiles
- Given appropriate historical data, have the ability to predict: “Under what conditions does vegetation not recover within a certain time frame.”
- Do global scale analysis in parallel

VESsel GENeration (VESGEN) Analysis

Patricia Parsons-Wingerter, PhD, NASA Chief Innovator/POC
NASA Ames 2016 Innovation Fund Award, Chief Technologist's Office

- VESGEN 2D maps and quantifies vascular remodeling for a wide variety of quasi-2D vascularized biomedical tissue applications.
- Working on transforming to VESGEN 3D, in line with most vascularized organs and tissues in humans and vertebrates.
- Vascular-dependent diseases include cancer, diabetes, coronary vessel disease, and major astronaut health challenges in the space microgravity and radiation environments, especially for long-duration missions.
- One key component is binarization: conversion of grayscale images to black/white vascular branching patterns.
  - Takes 10-25 hours of human effort.
  - Exploring pattern recognition, matching filtering, vessel tracking/tracing, mathematical morphology, multiscale approaches, and model based algorithms.

OTSU Thresholding



OTSU vs. Adaptive Thresholding

Future Work

- Work in progress: exploring more preprocessing and post-processing techniques
- Each step of preprocessing and postprocessing has some input parameters
  - The result is sensitive to this parameters
  - We aim to make the parameter selection either automated (machine learning) or semi-automated (user can choose the right parameter)
- Machine Learning to learn the binarization
  - Given the manual labels, perform supervised or semi-supervised learning
  - Each pixel and its class label (foreground or background) is the training example

How do we get the Word Out?

DASHlink

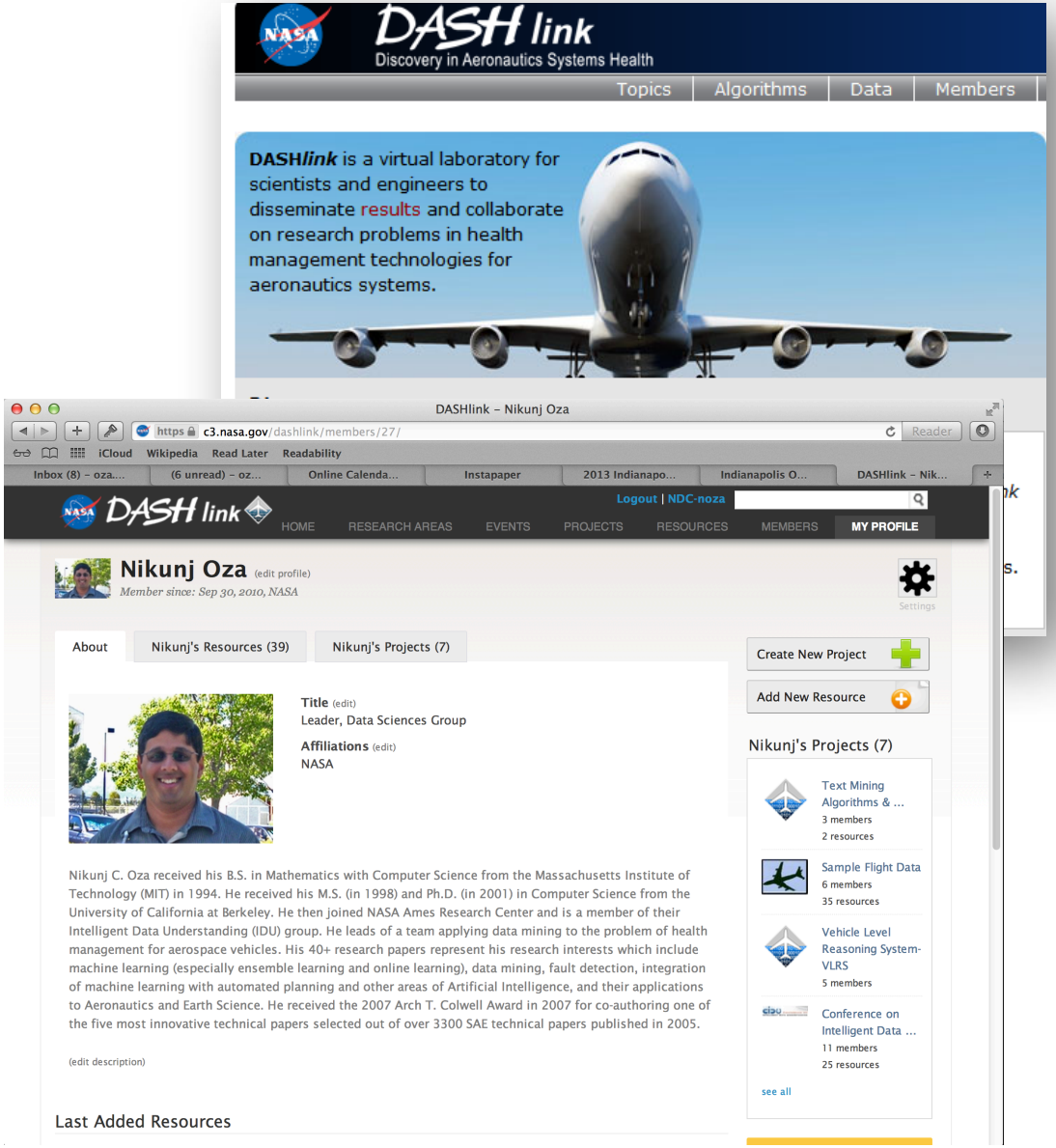
disseminate. collaborate. innovate.
https://dashlink.ndc.nasa.gov/

DASHlink is a collaborative website designed to promote:

- Sustainability
- Reproducibility
- Dissemination
- Community building

Users can create profiles

- Share papers, upload and download open source algorithms
- Find NASA data sets.




National Aeronautics and Space Adr
NASA
NASA Ames Data Sciences Group
Nikunj C. Oza, Ph.D.
Leader, Data Sciences Group
nikunj.c.oza@nasa.gov
www.nasa.gov
•21