

# *Machine Learning for NASA Earth Science Data Systems*

Manil Maskey, Ph.D.

Lead, NASA SMD Strategic Data Management Working Group on AI/ML, NASA HQ

Senior Research Scientist, Earth Science Data Systems, NASA HQ

Deputy Manager, NASA IMPACT, MSFC



# NASA Science Mission Directorate (SMD) Artificial Intelligence and Machine Learning (AI/ML) Initiatives

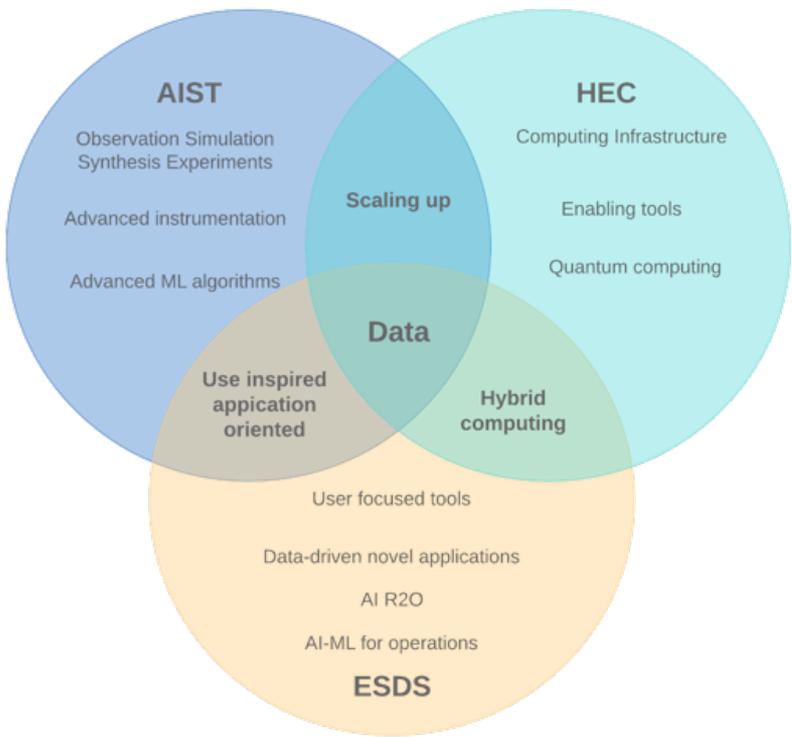
SMD's Strategy for Data Management and Computing for Groundbreaking Science 2019-2024 Report identified that AI/ML has yet to be fully appreciated and understood by SMD and science disciplines

## Activities:

- Identify areas of natural collaborations on AI/ML across SMD
- Conduct expert workshop on AI for science
- Explore industry partnership to transform data systems and leverage open science data
- Develop a roadmap to leverage **large volumes of data and computation** to accelerate AI/ML across NASA science



# ESD Programs Supporting AI/ML



*ESD has invested in AI-ML across the Division*

Algorithms and Hardware infrastructure

Data is key element of AI systems

Cloud Computing

Catalogs and Standardization

*R&A – individual PIs may use AI-ML techniques tailored to disciplines*



# ESDS AI/ML strategic goals

Goal 1: Augment scientific data stewardship processes

Goal 2: Maximize information and knowledge discovery capabilities

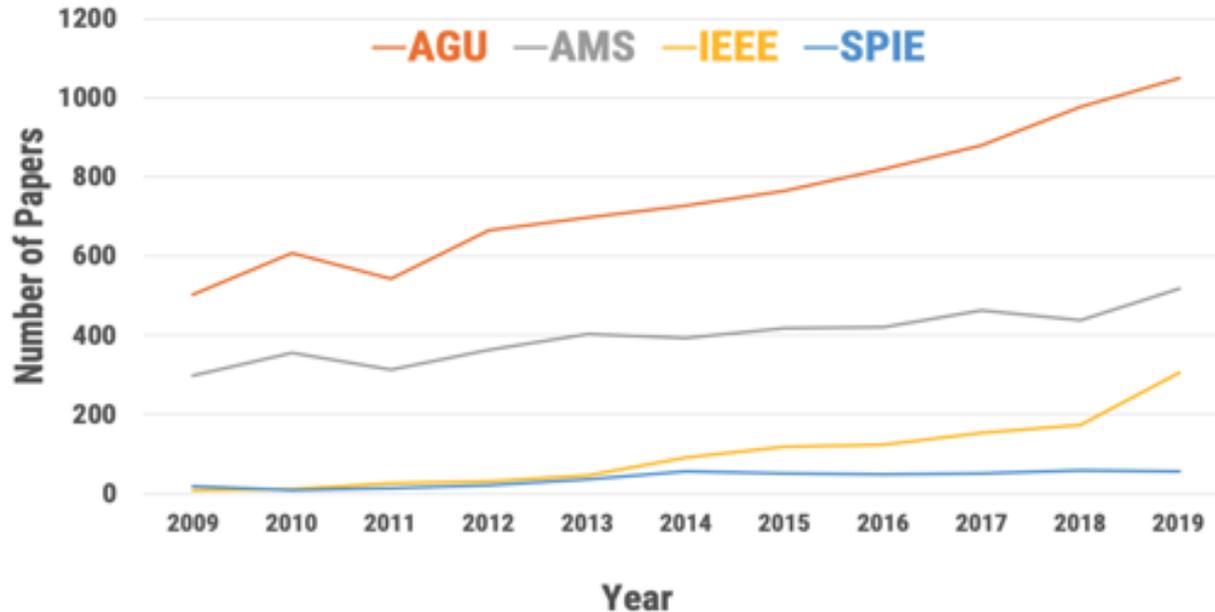
Goal 3: Enable sharing, and interoperability of Earth science AI training data and models

Goal 4: Increase engagement with commercial, academic, other agencies, and international partners

Goal 5: Foster AI expertise within the program



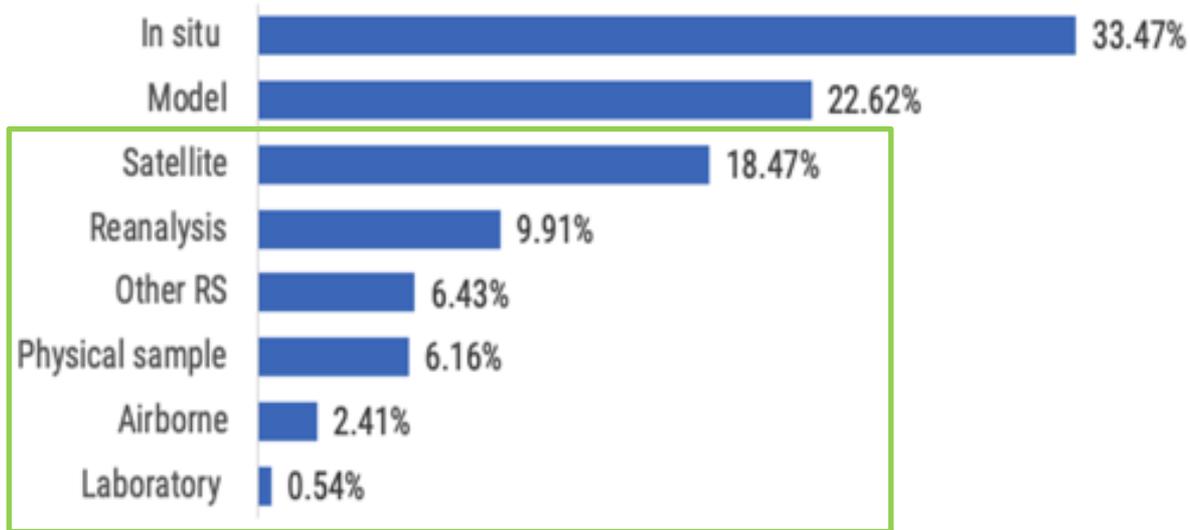
# AI/ML in Earth Science



Rapid adoption of AI/ML by Earth science researchers



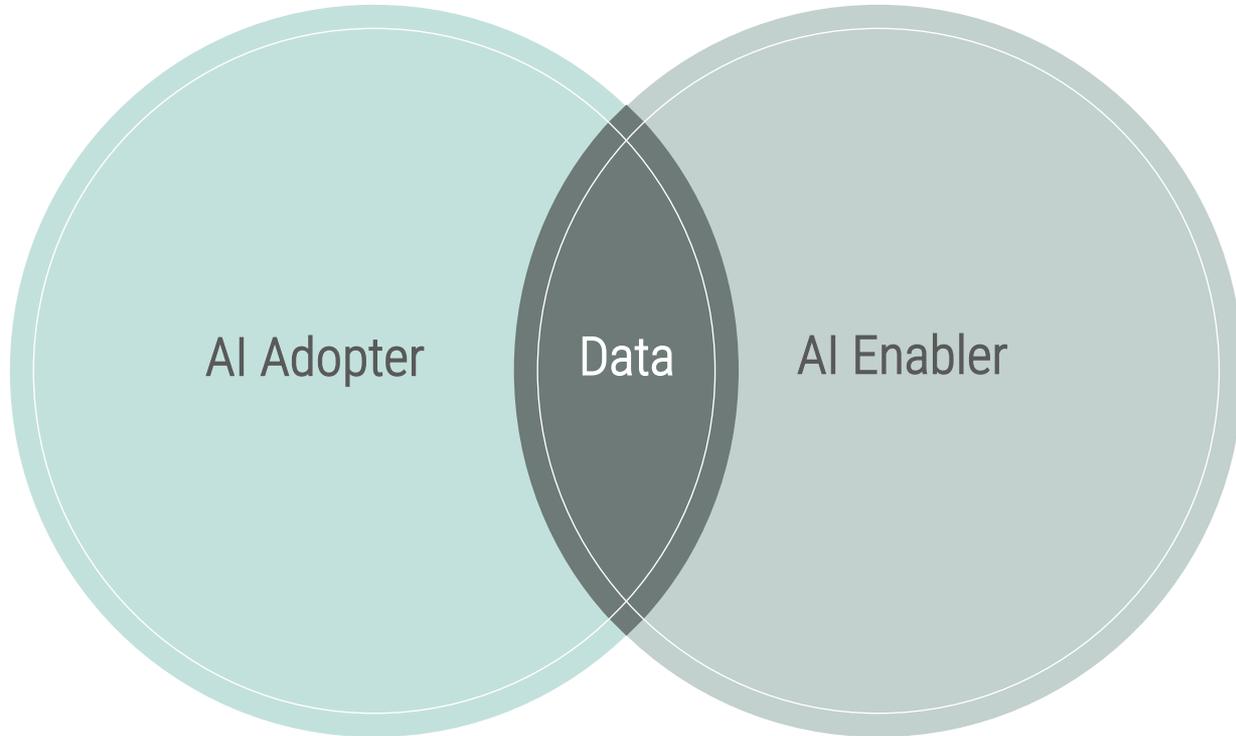
## Types of data used in Earth Science supervised machine learning papers



ESDS archive is underutilized by AI applications, and new data curation activities are needed to address the lack of training data



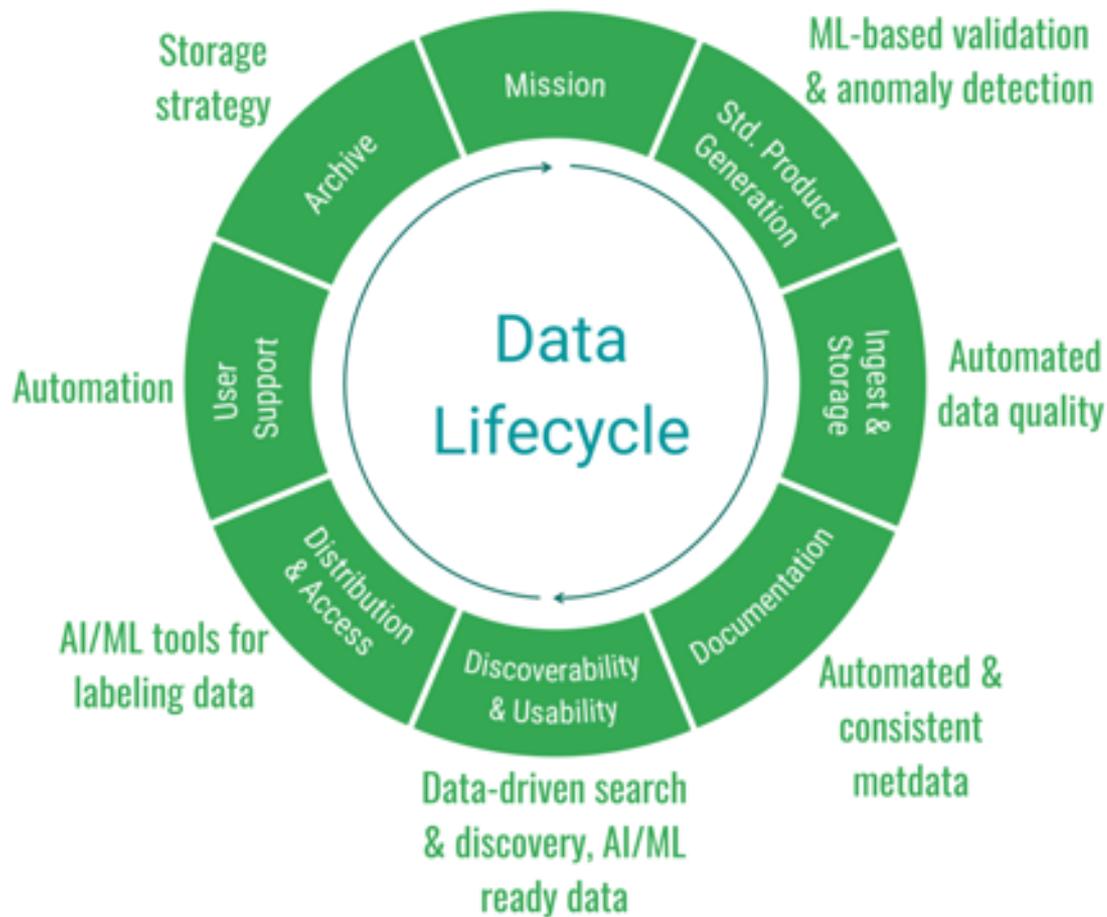
# AI and Data Systems



# Data systems as adopter of AI

- Improve efficiency of NASA's data systems operations
- Increase opportunity for researchers and commercial users to access/process PBs of data quickly without the need for data management
- Ensure users find right data for their problem
- Minimize user burden to access data
- Enable users to extract new knowledge/information from archives

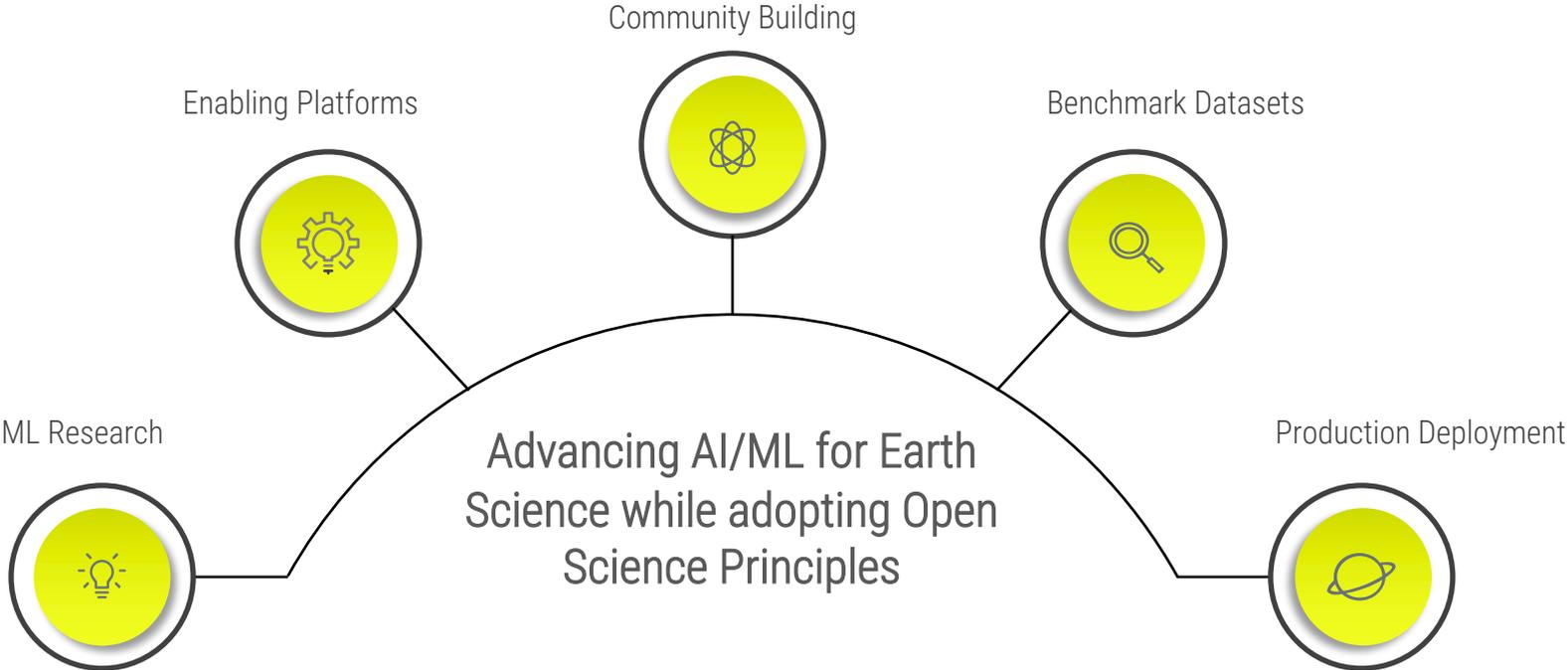




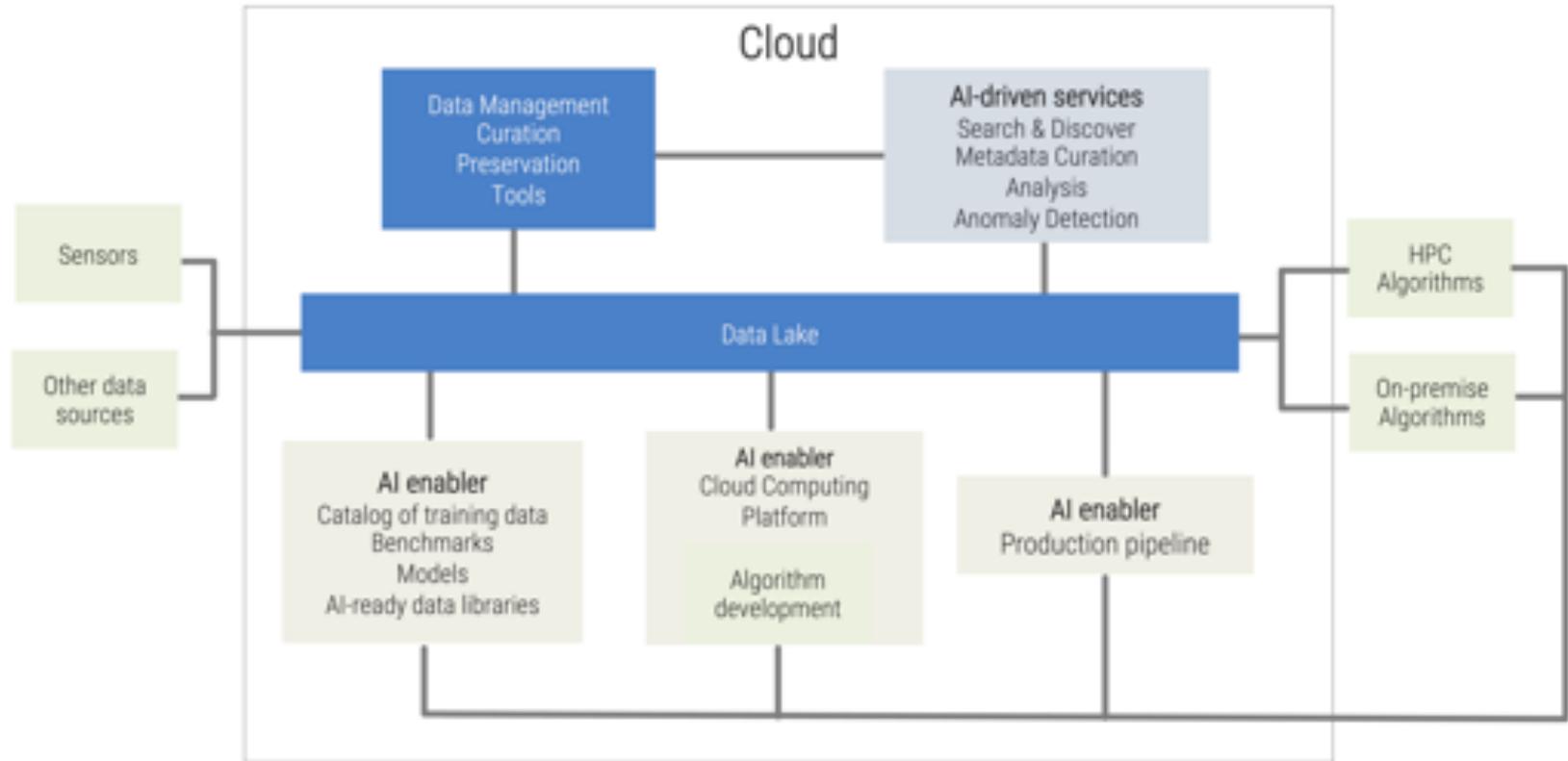
## AI for full data lifecycle



# Data systems as enabler of AI



# Data systems AI architecture design



# Current projects/activities

## Maximize information and knowledge discovery capabilities – computer vision

Novel search of archived data

## Augment data stewardship processes – NLP/Knowledge Graph

Keyword assignment

Metadata curation

## Accelerate AI for Earth science – AI-ready data

Labeled dataset generation/stewardship

Data science challenges

Labeling tools

## Enabling Platforms

Cloud/HPC

Deployment pipeline

## Partnerships

Industry/Academia involvement

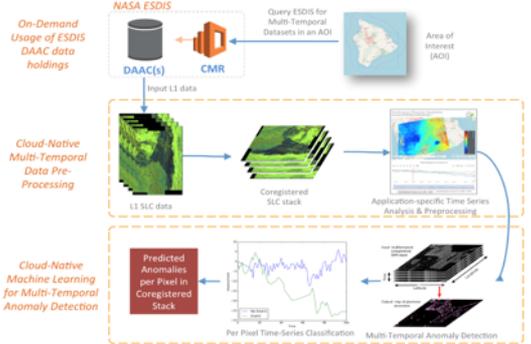
Workshops/Hackathons



# Ongoing AI/ML activities

## ACCESS

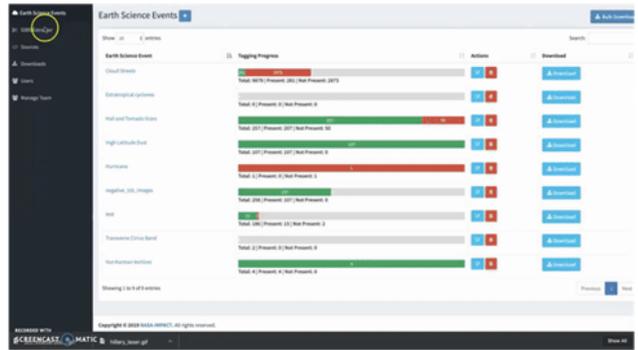
2017: Multi-temporal anomaly detection for SAR  
 2019: Several AI/ML focused projects



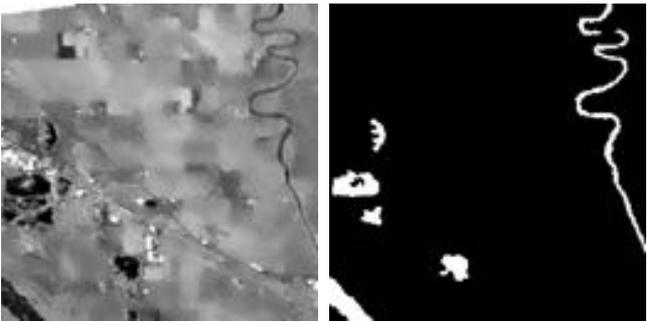
## Workshops



## Labeling tool

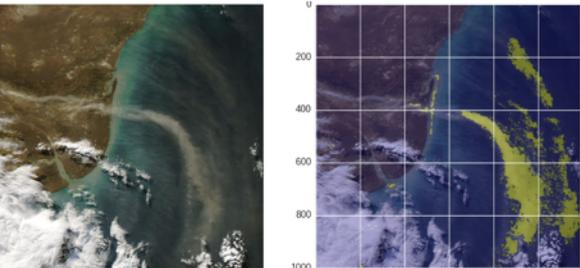


## Training dataset generation



Flood extent dataset on Sentinel 1

## Rapid prototypes



High latitude dust detection

## Data science challenges

Wind-dependent Variables: Predict Wind Speeds of Tropical Storms

WINSTED BY RANDALL SARTON FOUNDATION

User or team	Best points (1)	Score (0)	Timestamp	Trend (acc 10)	# Entries
ML4	1	6.2558	2021-01-31 10:56:04		105
ML4	1	6.4173	2021-01-28 14:21:22		81
ML4	1	6.4561	2021-01-31 18:25:53		59
ML4	1	6.4818	2021-02-02 17:25:06		43
ML4	1	6.6134	2021-01-30 21:52:41		126
ML4	1	6.6141	2021-01-28 08:55:51		116
ML4	1	6.6232	2021-02-02 19:37:17		67
ML4	1	6.7598	2021-02-02 22:46:56		111
ML4	1	6.8469	2021-02-01 17:56:21		27
ML4	1	6.8585	2021-02-01 00:13:58		56
ML4	1	7.0325	2021-02-02 23:26:32		142
ML4	1	7.3024	2021-01-28 15:26:58		24

Throughout a tropical cyclone, humanitarian response efforts hinge on accurate storm intensity estimates. Using satellite images processed by the NASA Earth Foundation and the NASA IMPACT team, can you estimate the wind speeds of storms at different times?

**Quick Facts**

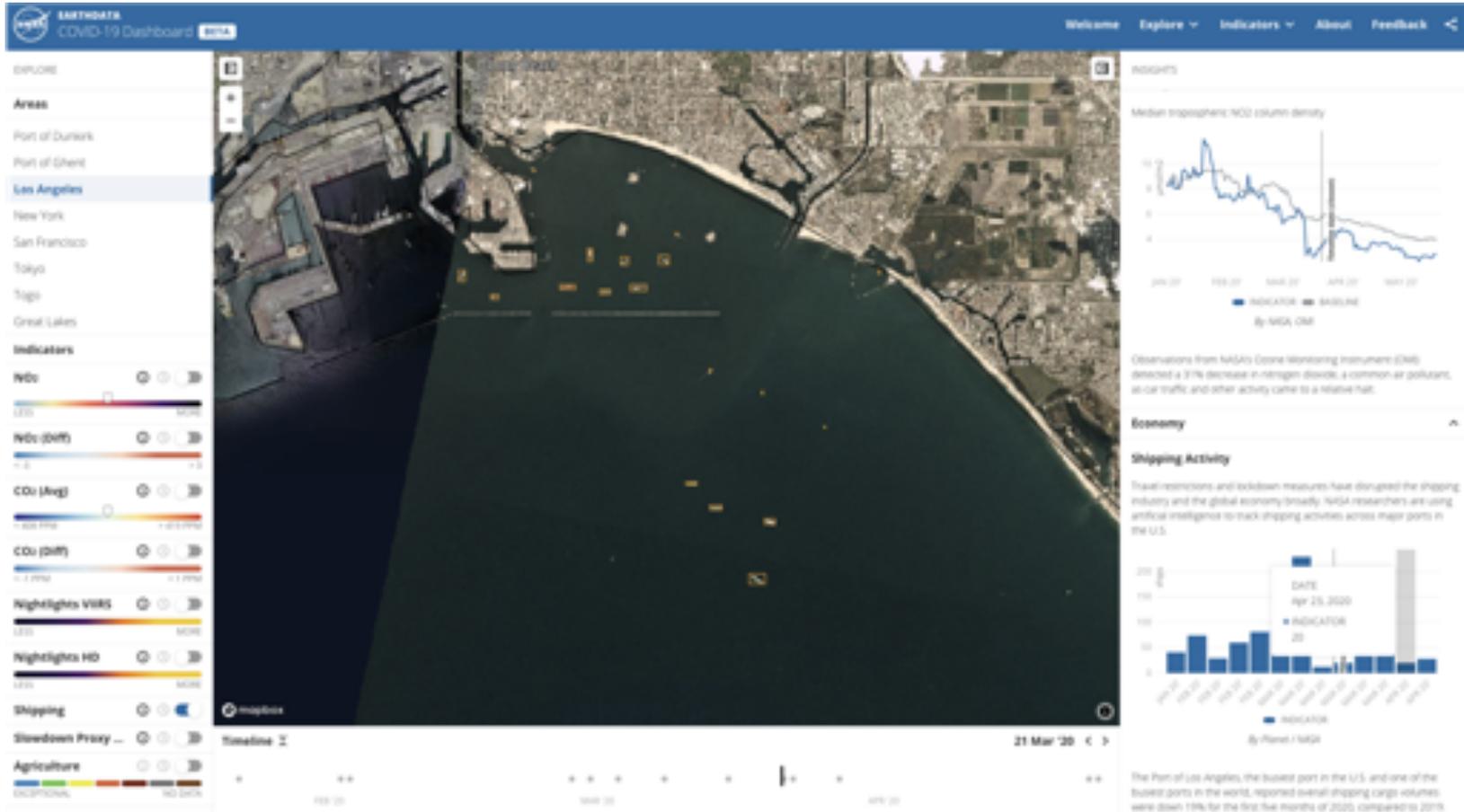
- PARTICIPANTS: 733
- NO. OF ENTRIES: 2,755
- PRIZE: \$13,000
- WINNER: ML4 VECKOZ 1ST PLACE

LEADERBOARD RESULTS

OFFICIAL RULES



# Leveraging AI+CSDAP+Prototypes+Cloud for COVID-19 Dashboard



# Resources

<a href="https://labeler.nasa-impact.net/">Image labeler</a>	<a href="https://labeler.nasa-impact.net/">https://labeler.nasa-impact.net/</a>
<a href="http://phenomena.surge.sh/">Phenomena portal</a>	<a href="http://phenomena.surge.sh/">http://phenomena.surge.sh/</a>
<a href="http://hurricane.dsig.net/">Hurricane intensity estimation portal</a>	<a href="http://hurricane.dsig.net/">http://hurricane.dsig.net/</a>
<a href="https://earthdata.nasa.gov/covid19/">COVID-19 dashboard</a>	<a href="https://earthdata.nasa.gov/covid19/">https://earthdata.nasa.gov/covid19/</a>
<a href="https://impact.earthdata.nasa.gov/">IMPACT website</a>	<a href="https://impact.earthdata.nasa.gov/">https://impact.earthdata.nasa.gov/</a>
<a href="https://earthdata.nasa.gov/esds/competitive-programs/access">ACCESS 2019 projects</a>	<a href="https://earthdata.nasa.gov/esds/competitive-programs/access">https://earthdata.nasa.gov/esds/competitive-programs/access</a>

Thank you.

Manil Maskey  
manil.maskey@nasa.gov



Backup slides



Maximize information and knowledge  
discovery capabilities



# Why?

Increasing Earth science data archives require non-traditional approaches to data management

Data driven technologies to provide advanced search capabilities

Machine learning-based approach - an enabling data driven technology to provide automated detection of Earth science events from image archives

Catalog of events can provide a novel way to explore large archives of data

Discover and explore Earth science data archives around events using machine learning (ML) techniques





FEATURES

Chlorophyll  
Chlorophyll (SeaWiFS) (Global) (SeaWiFS)

Coastlines / Borders / Roads  
Coastlines (Global) (Global)

Global Precipitation (IMERG)  
Global Precipitation (IMERG)

Global Temperature (M2500)  
Global Temperature (M2500)

Sea and Thermal Anomaly (SST)  
Sea and Thermal Anomaly (SST)

Sea Level (TOPEX/Poseidon)  
Sea Level (TOPEX/Poseidon)

Sea Level (TOPEX/Poseidon)

Sea Level (TOPEX/Poseidon)

Sea Level (TOPEX/Poseidon)

Sea Level (TOPEX/Poseidon)

Sea Level (TOPEX/Poseidon)

Sea Level (TOPEX/Poseidon)

Sea Level (TOPEX/Poseidon)

Sea Level (TOPEX/Poseidon)

Sea Level (TOPEX/Poseidon)

Sea Level (TOPEX/Poseidon)

Sea Level (TOPEX/Poseidon)

Sea Level (TOPEX/Poseidon)

Sea Level (TOPEX/Poseidon)

Sea Level (TOPEX/Poseidon)

Sea Level (TOPEX/Poseidon)

Sea Level (TOPEX/Poseidon)

Sea Level (TOPEX/Poseidon)

Sea Level (TOPEX/Poseidon)

Sea Level (TOPEX/Poseidon)

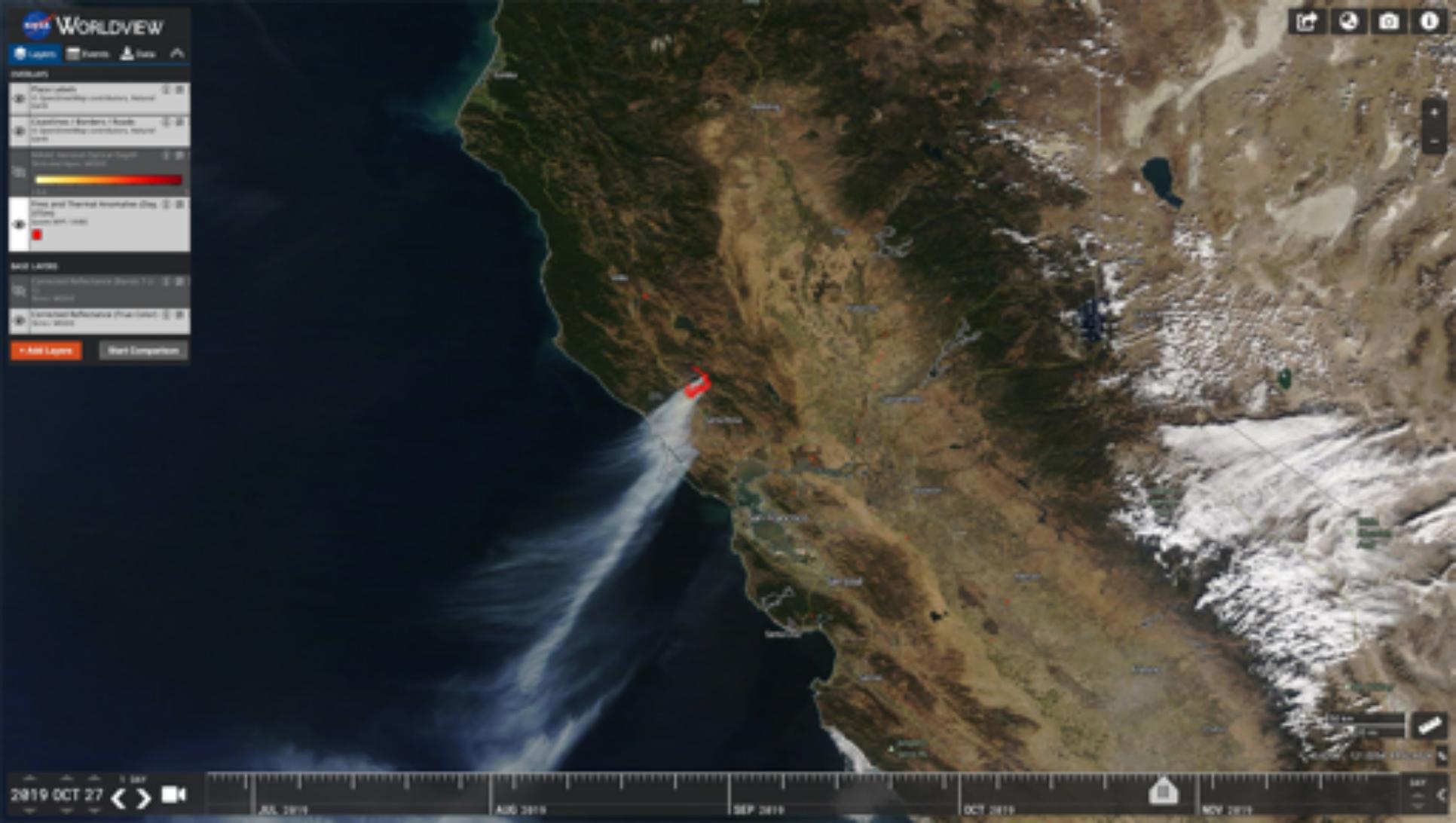
Sea Level (TOPEX/Poseidon)

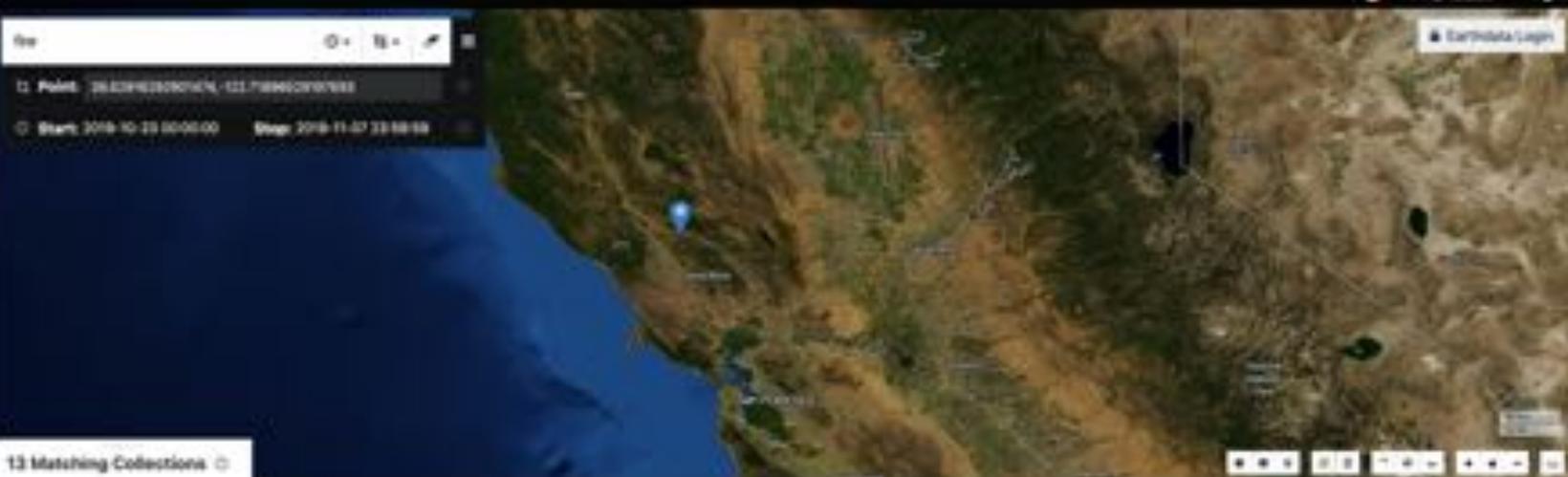
Sea Level (TOPEX/Poseidon)

Sea Level (TOPEX/Poseidon)

Sea Level (TOPEX/Poseidon)

Sea Level (TOPEX/Poseidon)





## 13 Matching Collections

 Sort by: Relevance  Only include collections with granules  Include non-EODSIS collections 

You will be redirected to your project to compare and download their data.


**Woods/Bass Thermal Anomalies/Pre 8-Day L2 Global Sea SW Grd V006**

8 Granules • 2000-02-18 ending • The Terra Moderate Resolution Imaging Spectroradiometer (MODIS) Thermal Anomalies and Pre 8-Day (MOT1442) Version 6 data are generated at 1-kilometer (km) spatial resolution as a Level 2 product. The MOT1442 gridded composite contains the maximum value of the individual five pixel values detected during the eight days of acquisition. The Science Dataset (SDS) layers include the file mask and pixel quality indicators. Improvements/Changes from Previous Versions: "

[View this collection's profile](#)

**Woods/Bass Thermal Anomalies/Pre Daily L2 Global Sea SW Grd V006**

8 Granules • 2000-02-18 ending • The Terra Moderate Resolution Imaging Spectroradiometer (MODIS) Thermal Anomalies and Pre Daily (MOT1443) Version 6 data are generated every eight days at 1-kilometer (km) spatial resolution as a Level 2 product. MOT1443 contains eight consecutive days of the data consistently packaged into a single file. The Science Dataset (SDS) layers include the file mask, pixel quality indicators, maximum the relative power (MaxRPP), and the position of the five pixel within the sea...

[View this collection's profile](#)

**Woods/Bass Thermal Anomalies/Pre Daily L2 Global Sea SW Grd V006**

8 Granules • 2000-07-04 ending • The Terra Moderate Resolution Imaging Spectroradiometer (MODIS) Thermal Anomalies and Pre Daily (MOT1443) Version 6 data are generated every eight days at 1-kilometer (km) spatial resolution as a Level 2 product. MOT1443 contains eight consecutive days of the data consistently packaged into a single file. The Science Dataset (SDS) layers include the file mask, pixel quality indicators, maximum the relative power (MaxRPP), and the position of the five pixel within the sea...

[View this collection's profile](#)

**Woods/Bass Thermal Anomalies/Pre 8-Day L2 Global Sea SW Grd V006**

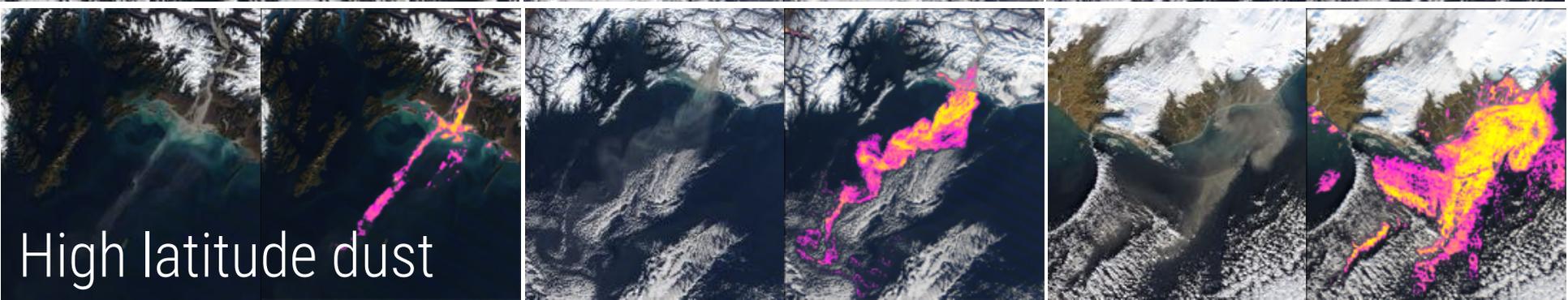
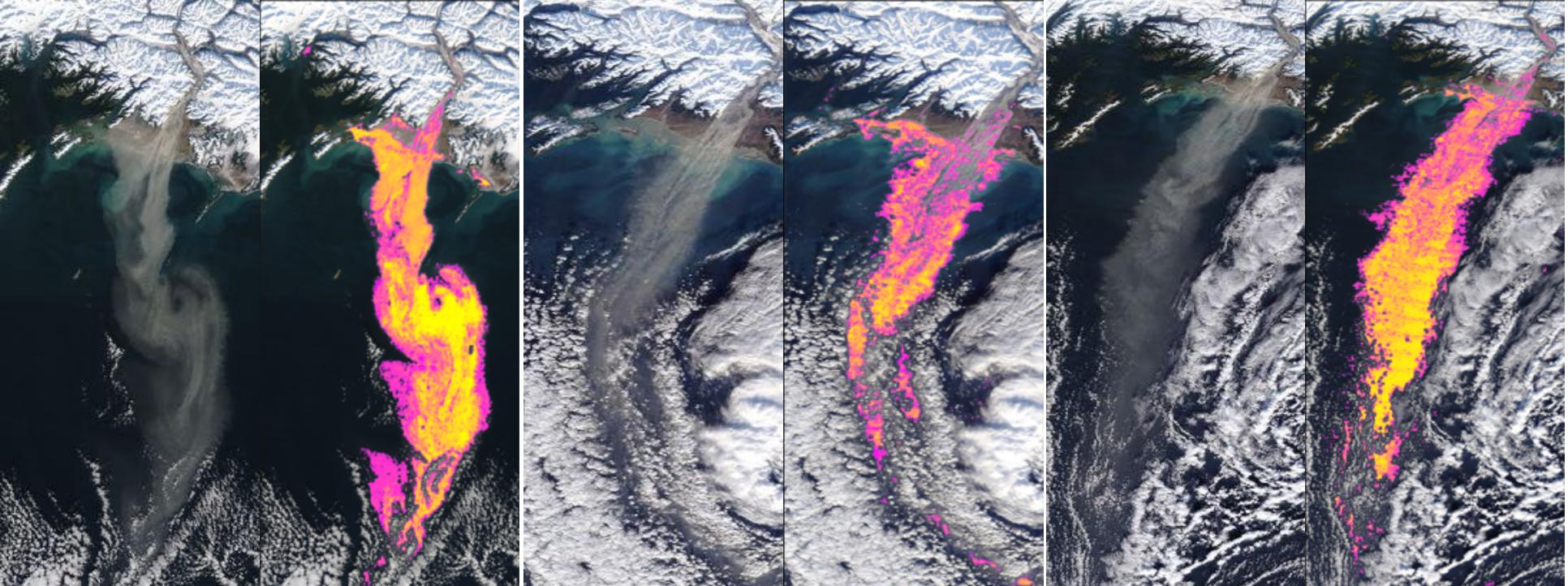
8 Granules • 2000-07-04 ending • The Terra Moderate Resolution Imaging Spectroradiometer (MODIS) Thermal Anomalies and Pre 8-Day (MOT1442) Version 6 data are generated at 1-kilometer (km) spatial resolution as a Level 2 product. The MOT1442 gridded composite contains maximum value of individual five pixel values detected during the eight days of acquisition. The Science Dataset (SDS) layers include the file mask and pixel quality indicators. Improvements/Changes from Previous Versions: "

[View this collection's profile](#)



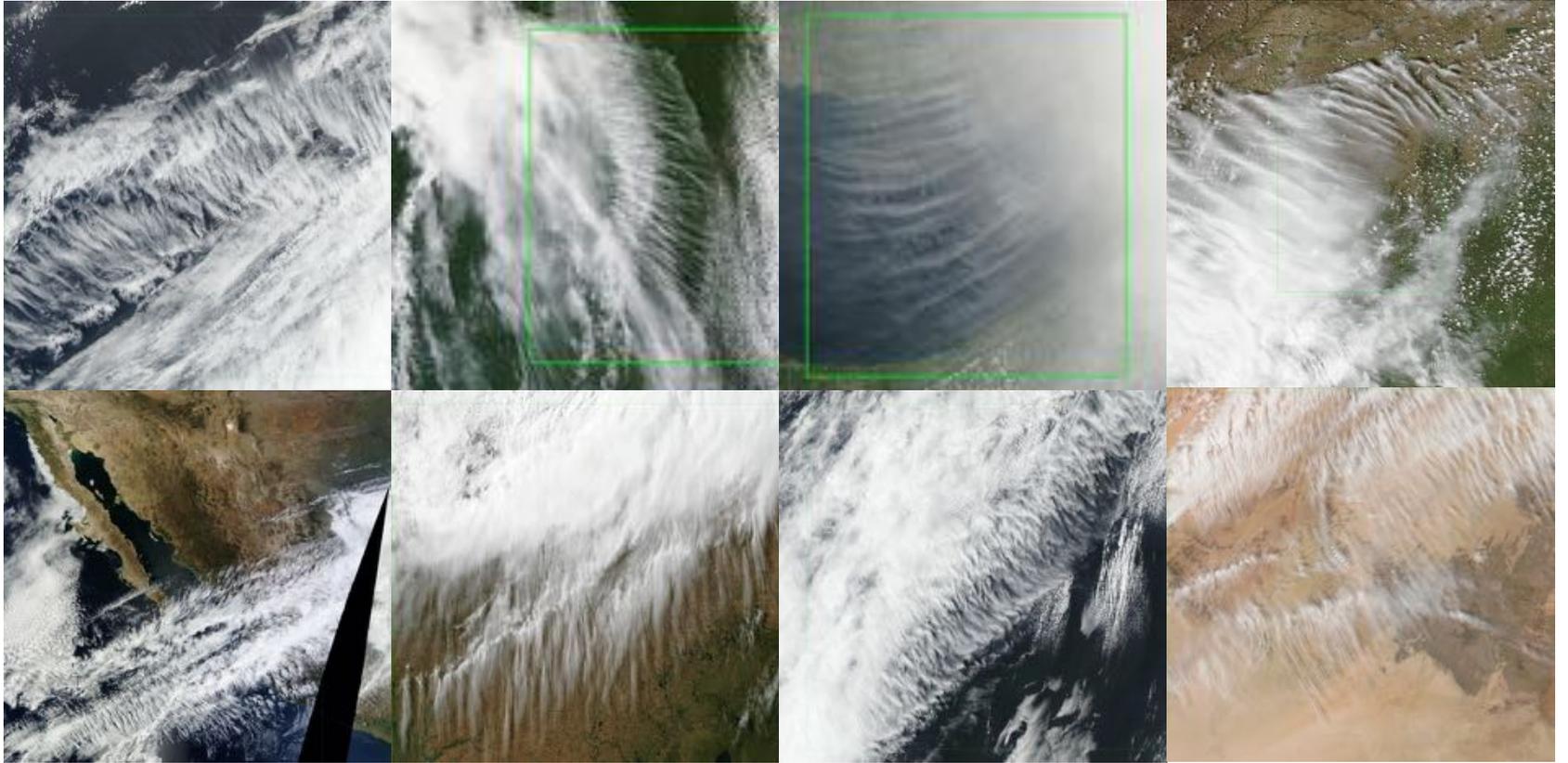

Demo Video





High latitude dust

# Transverse cirrus bands



# Welcome to the Phenomena Detection Portal

We are using machine learning for real-time detection of Earth science phenomena.

Types

**03**

so far

Detections

**98,627**

and counting

Confidence score

**89.61%**

on average

[Start exploring](#)

[Learn more](#)

[Demo Video](#)



# Augment data stewardship processes

## *Automated keyword assignment*



# Why?

Assigning science keywords is currently a manual process, which is prone to human error and inconsistencies.

Metadata managed across a network of multiple data centers (i.e. keywords not assigned by a central entity)

Keywords may be assigned by non-subject matter experts (SMEs)

Improve metadata quality

Provide objective and consistent approach to keyword assignment

LIS/OTD 2.5 Degree Low Resolution Annual Climatology Time Series (LRACTS) V2.3.2015

## CMR Dataset Title and Description

### Abstract

The LIS/OTD 2.5 Degree Low Resolution Annual Climatology Time Series (LRACTS) consists of gridded climatologies of total lightning flash rates seen by the spaceborne Optical Transient Detector (OTD) and Lightning Imaging Sensor (LIS). The long LIS (equatorward of about 38 degree) record makes the merged climatology most robust in the tropics and subtropics, while the high latitude data is entirely from OTD. The LRACTS dataset include annual flash rate time series data in MP4 format.

### DOI

10.5067/LIS/LIS-OTD/DATA306

### Science Keywords

EARTH SCIENCE Atmosphere Atmospheric Electricity Lightning

EARTH SCIENCE Atmosphere Weather Events Lightning



# Approach – build word embeddings

Journal Name	Date Published																		
	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	
Atmospheric Science Letters		5	21	34	27	22	42	39	33	44	34	30	69	62	99	65	67	32	
Earth and Space Science													1	24	28	25	42	88	
Earth's Future												13	26	24	32	31	80	36	
Estuaries and Coasts										46	37			1					
Geochemistry, Geophysics, Geosystems	86	202	286	373	355	343	328	314	264	283	287	286	154	186	167	64	29		
Geochimica et Cosmochimica Acta																22	34	14	
Geophysical Research Letters	1,100	1,436	1,550	1,696	1,700	1,515	1,509	1,390	1,099	1,294	1,026	1,114	1,266	1,369	1,491	1,390	1,507	1,338	
Global Biogeochemical Cycles	137	136	124	128	75	46	34	66	83	64	36	76	77	26	75	74	73	71	
Journal of Advances in Modeling Earth Systems							6	3	3	3	25	25	49	88	56	113	125	134	
Journal of Geophysical Research							36				20								
Journal of Geophysical Research: Atmospheres	111	1,296	786	756	264	565	122	137	159	340	280	36	111	284	411	256	264	128	
Journal of Geophysical Research: Biogeosciences				23	29	140	111	117	146	185	130	57	100	133	138	137	45	24	
Journal of Geophysical Research: Earth Surface		52	47	91	84	145	130	113	134	137	141	117	67	67	93	58	44	30	
Journal of Geophysical Research: Oceans	254	357	325	314	317	303	434	323	367	501	418	336	349	330	325	306	352	264	
Journal of Geophysical Research: Planets	137	279	175	124	167	150	135	130	147	171	177	207	89	88	75	111	147	67	
Journal of Geophysical Research: Solid Earth	345	603	465	319	377	435	436	345	509	470	376	297	315	367	354	308	326	36	
Journal of Geophysical Research: Space Physics	424	543	436	525	475	447	505	533	256	271	484	490	503	543	530	547	466	446	
Meteosat									7	47	60	26	76	46	61	67	76	1	2
Palaeogeography	65	109	96	62	62	41	36	43	64	59	55	45	59	26	54	30			
Palaeogeography and Palaeoclimatology																	24	26	
Quarterly Journal of the Royal Meteorological Society									6	105		203	166	157	170				
Radio Science	104	137	146	114	94	122	91	206	10	136		79	57	63	79	100	91	94	51
Reviews of Geophysics	3	23	12	11	6	14					13	12	14	22	36	22	16	17	
Space Weather		16	37	51	48	47	44	46	30	55	43	66	53	67	48	88	125	58	
Tectonics	33	45	78	88	58	73	66	60	73	67	70	58	74	83	99	116	43	58	
Water Resources Research	302	296	319	317	329	364	314	254	350	401	406	447	412	365	404	317	413	446	

**88,410**  
documents

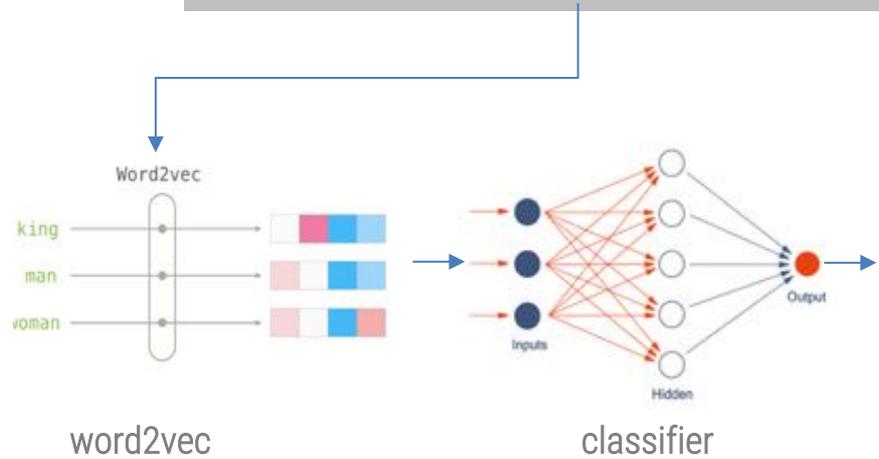
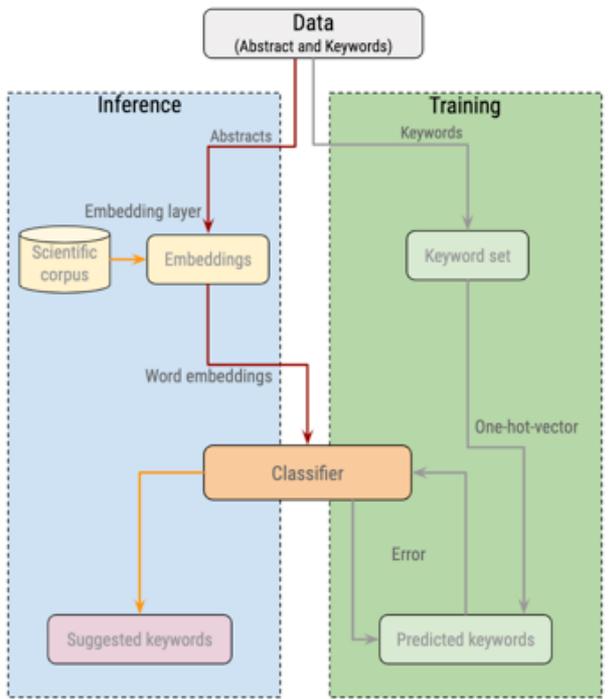
**530 million**  
words

**5.5 million**  
unique words



# Automated keyword assignment

Version 7.3 is the current version of the data set. Version 3.5 is no longer available and has been superseded by Version 7.3. This data set is currently provided by the OCO (Orbiting Carbon Observatory) Project. In expectation of the OCO-2 launch, the algorithm was developed by the Atmospheric CO2 Observations from Space (ACOS) Task as a preparatory project, using GOSAT TANSO-FTS spectra. After the OCO-2 launch, "ACOS" data are still produced and improved, using approaches applied to the OCO-2 spectra. The "ACOS" data set contains Carbon Dioxide (CO2) column averaged dry air mole fraction for all soundings for which retrieval was attempted. These are the highest-level products made available by the OCO Project, using TANSO-FTS spectral radiances, and algorithm build version 7.3. The GOSAT team at JAXA produces GOSAT TANSO-FTS Level 1B



Predicted Keyword	Score
carbon dioxide	0.4513424
land use/land cover classification	0.3825603
terrain elevation	0.1924277
barometric altitude	0.18085223
carbon and hydrocarbon compounds	0.07634798



## Deep Learning-based Hurricane Intensity Estimator

Applying machine learning to objectively estimate tropical cyclone intensity.

Explore

or

Read more



Don't show again

Demo Video