A collective agenda on AI for Earth sciences

Manil Maskey, Ph.D.

AI for Good
February 16, 2022

What I hope to do today

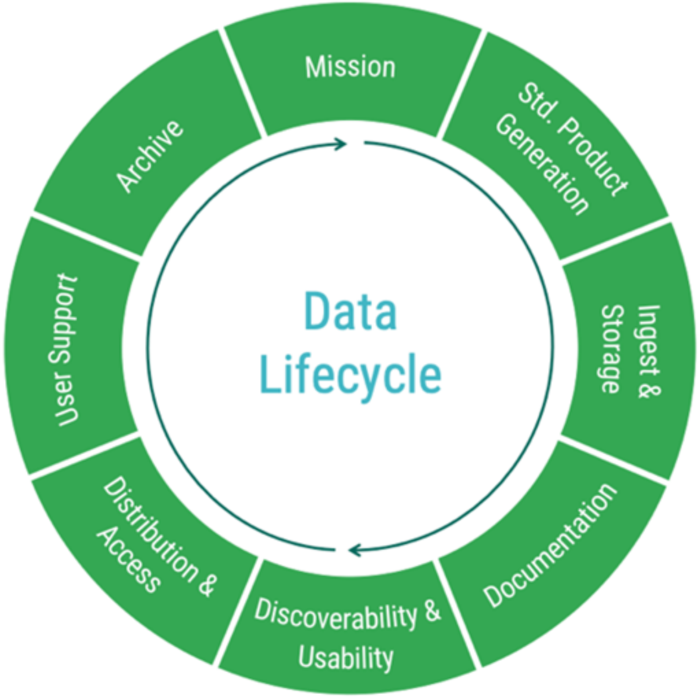
Data systems perspective on AI for Earth science

- AI for core data systems services
  - Search
  - Knowledge discovery
- Enabling AI to advance Earth science
  - Data (labeled training data) is the proprietary differentiator
  - Transitioning AI models to Production
  - Citizen science

NASA's Earth Science Data Systems Program

Single largest repository of Earth Science Data

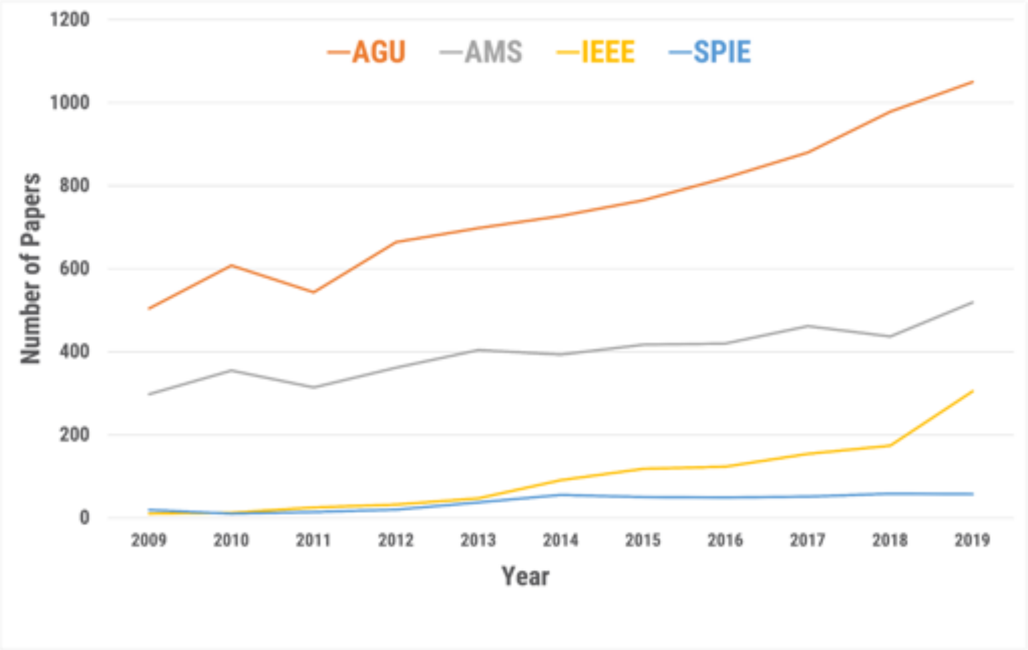
Manages NASA's Earth science data through the entire data life cycle



AGU topic trend

Number of journal articles per keyword by year

Publication trend - ML in Earth sciences



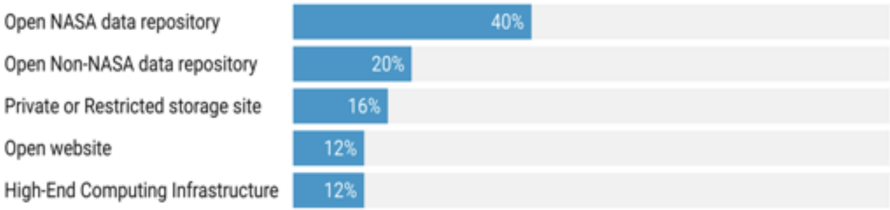
Rapid adoption of AI/ML by Earth science researchers

Virts et al. (2020)

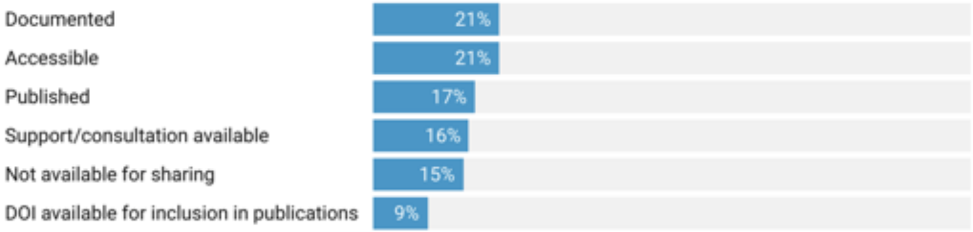
Maskey et al. (2020)

NASA Science survey - AI and data

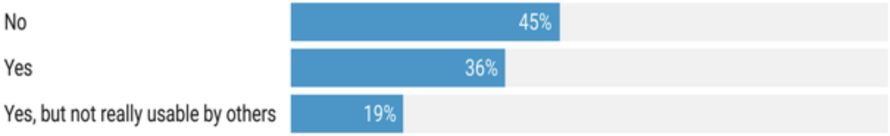
Source of data used



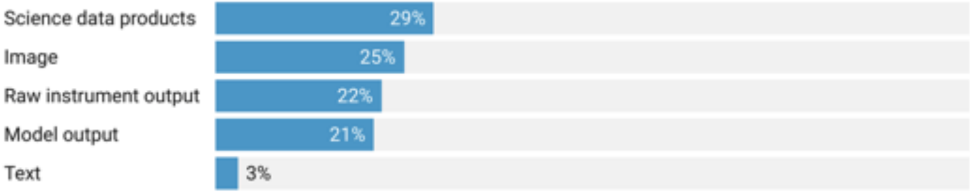
How re-usable is your training data?



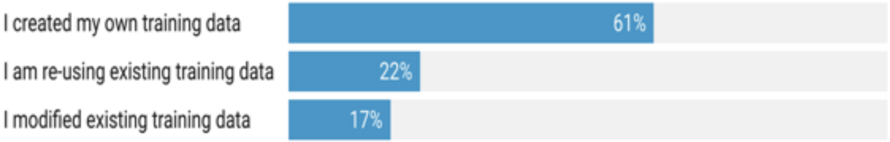
Is there a catalog of training data for your use?



What type of data do you use for AI?



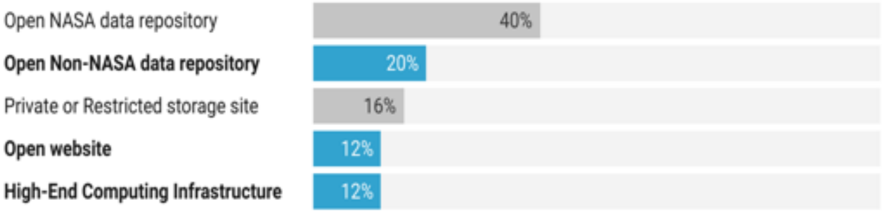
How did you construct training data?



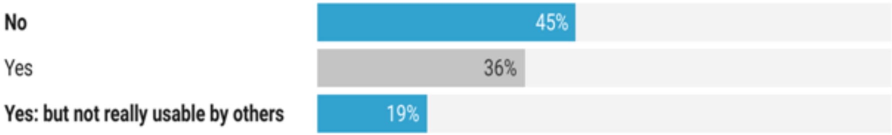
Amount of effort required to prepare data for AI?

NASA Science survey - AI and data

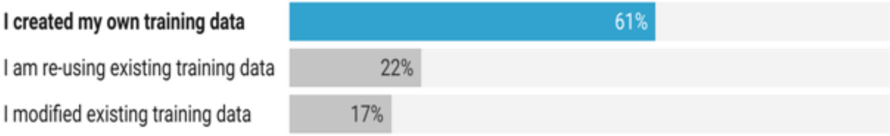
Source of data used



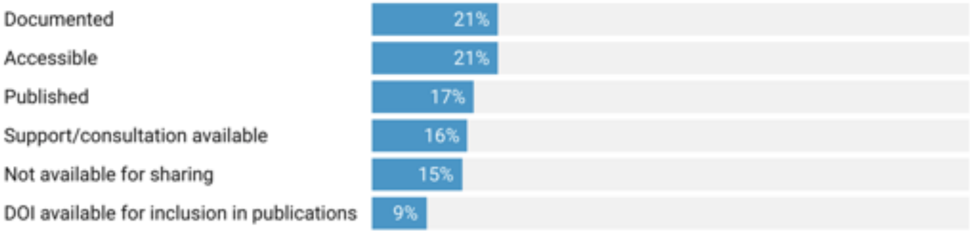
Is there a catalog of training data for your use?



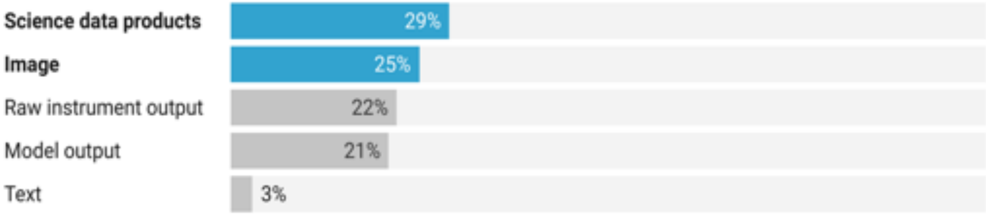
How did you construct training data?



How re-usable is your training data?



What type of data do you use for AI?



Amount of effort required to prepare data for AI?

Maximizing Knowledge Discovery

Why?

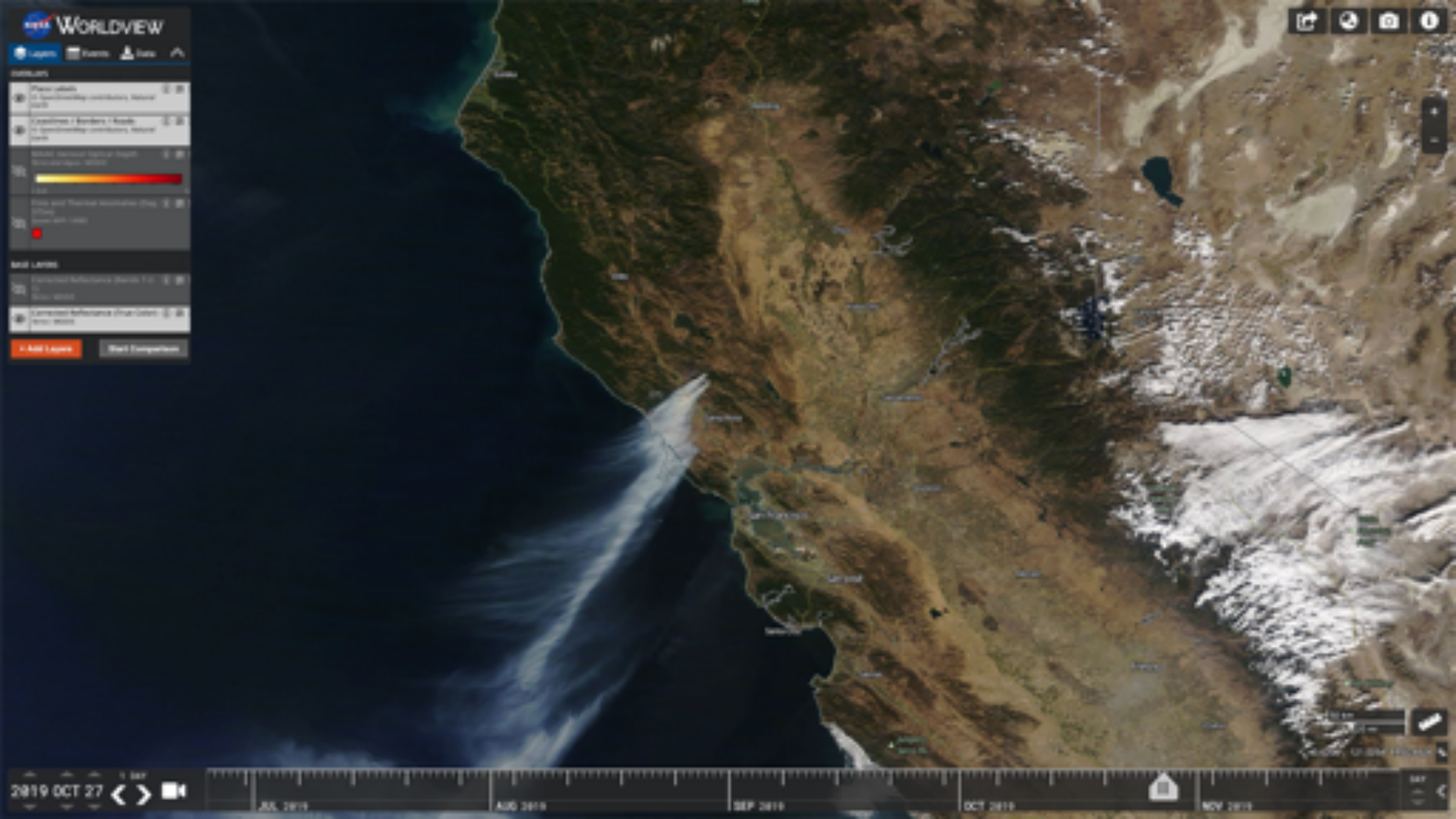
Increasing Earth science data archives require non-traditional approaches to data management

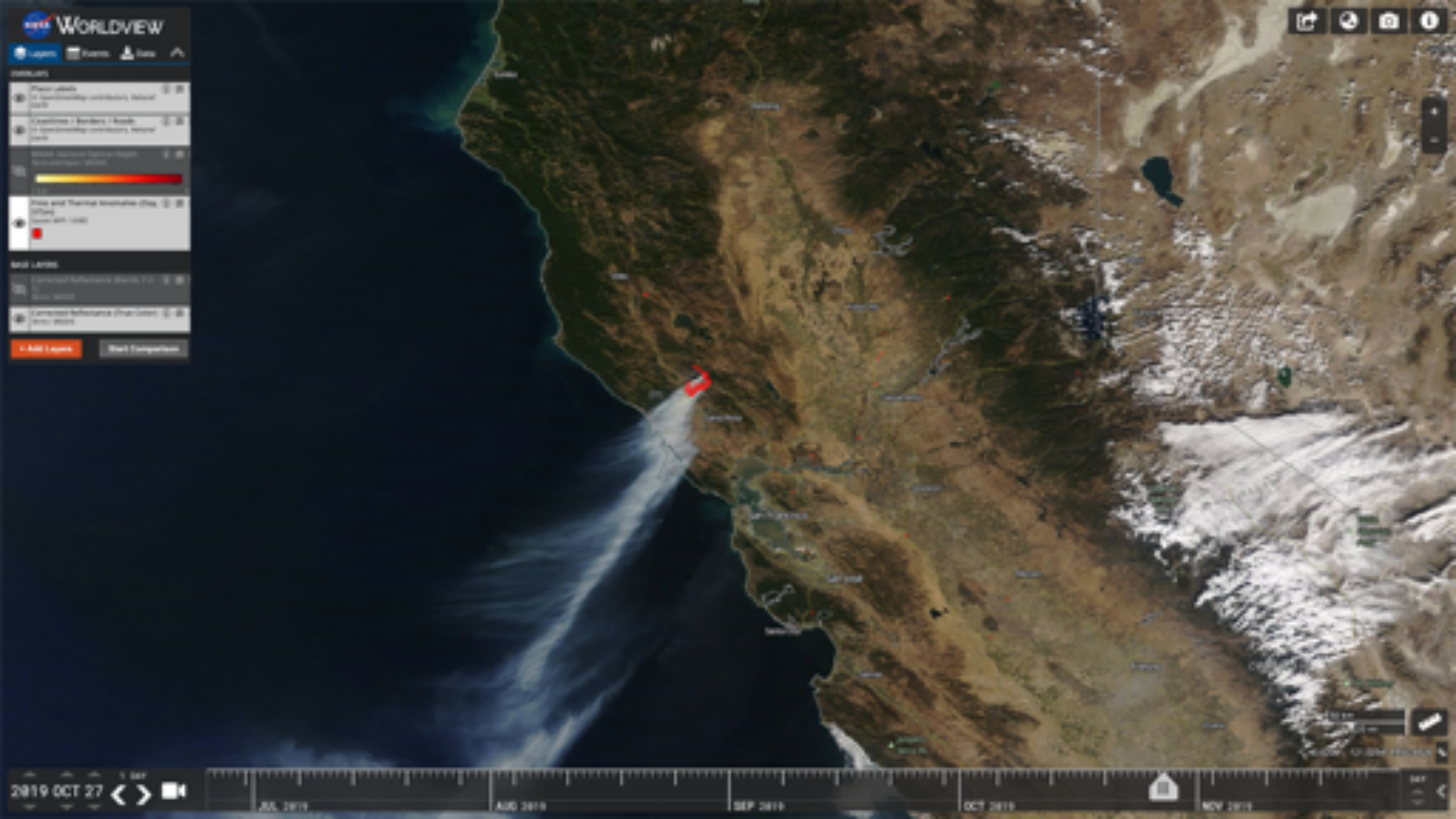
Data driven technologies (AI) to provide advanced search capabilities

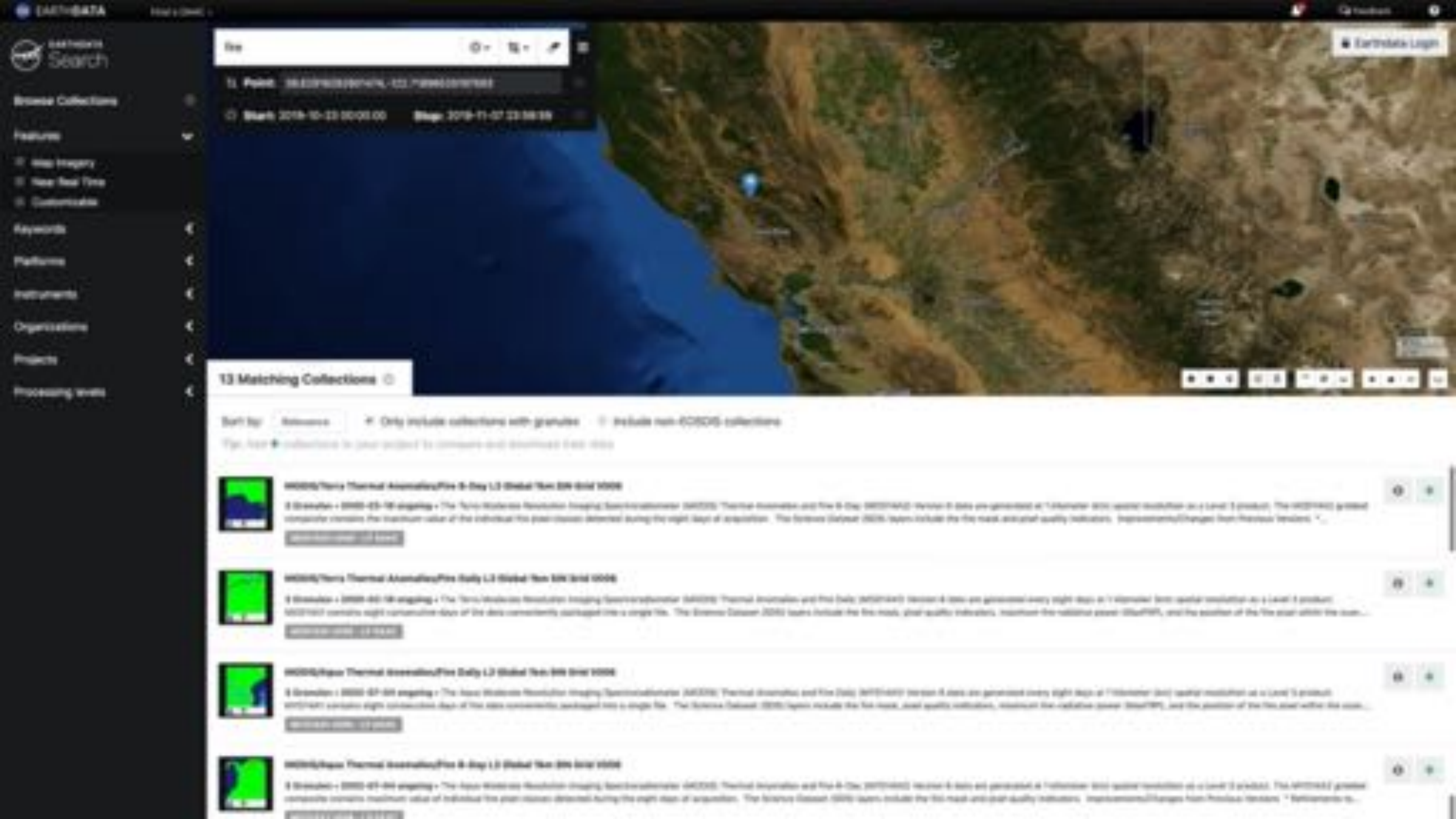
Machine learning-based approach - provide automated detection of Earth science events from image archives

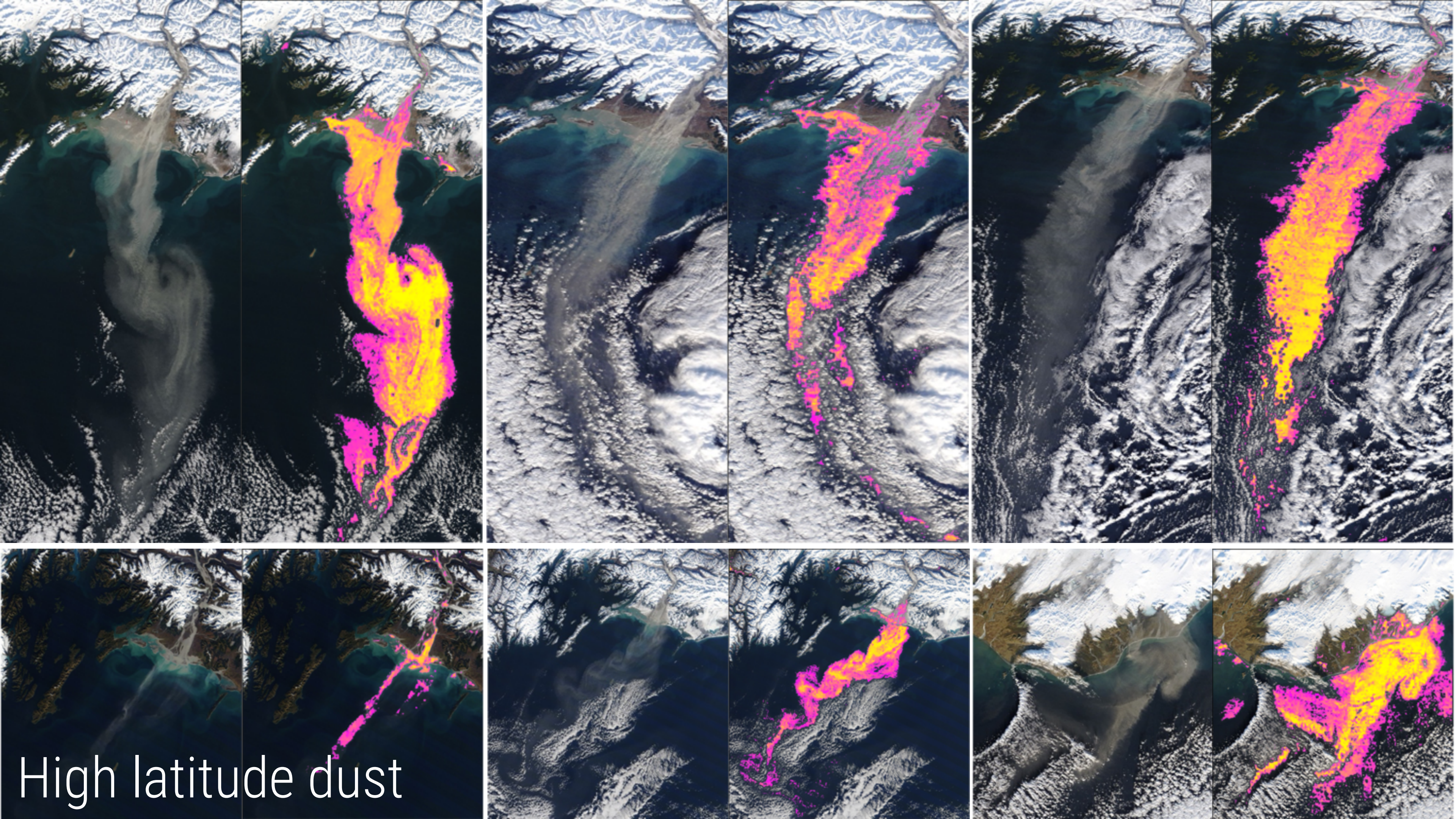
Catalog of events can provide a novel way to explore large archives of data

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

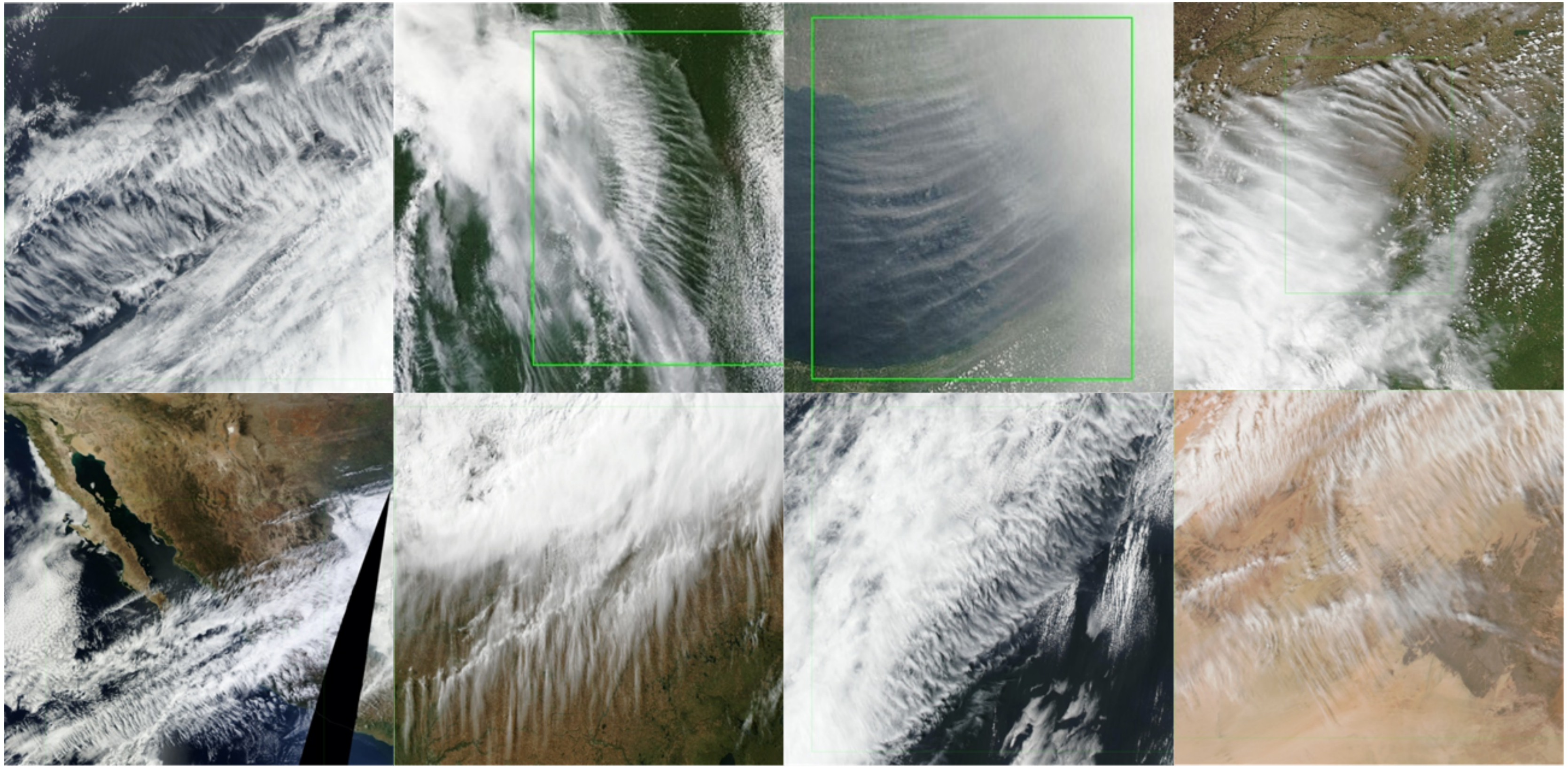

WORLDVIEW
2019 OCT 27
JUL 2019
AUG 2019
SEP 2019
OCT 2019
NOV 2019


WORLDVIEW
2019 OCT 27
JUL 2019
AUG 2019
SEP 2019
OCT 2019
NOV 2019


EARTHDATA
Earthdata Search
fire
Point:
Start:
Stop:
Earthdata Login
Browse Collections
Features
Map Imagery
Near Real Time
Customizable
Keywords
Platforms
Instruments
Organizations
Projects
Processing levels
13 Matching Collections
Sort by:
Relevance
Only include collections with granules
Include non-EOSDIS collections


High latitude dust

Transverse cirrus bands



Phenomena Detection Portal

Welcome | Explore | Phenomena | Methodology | About

Welcome to the

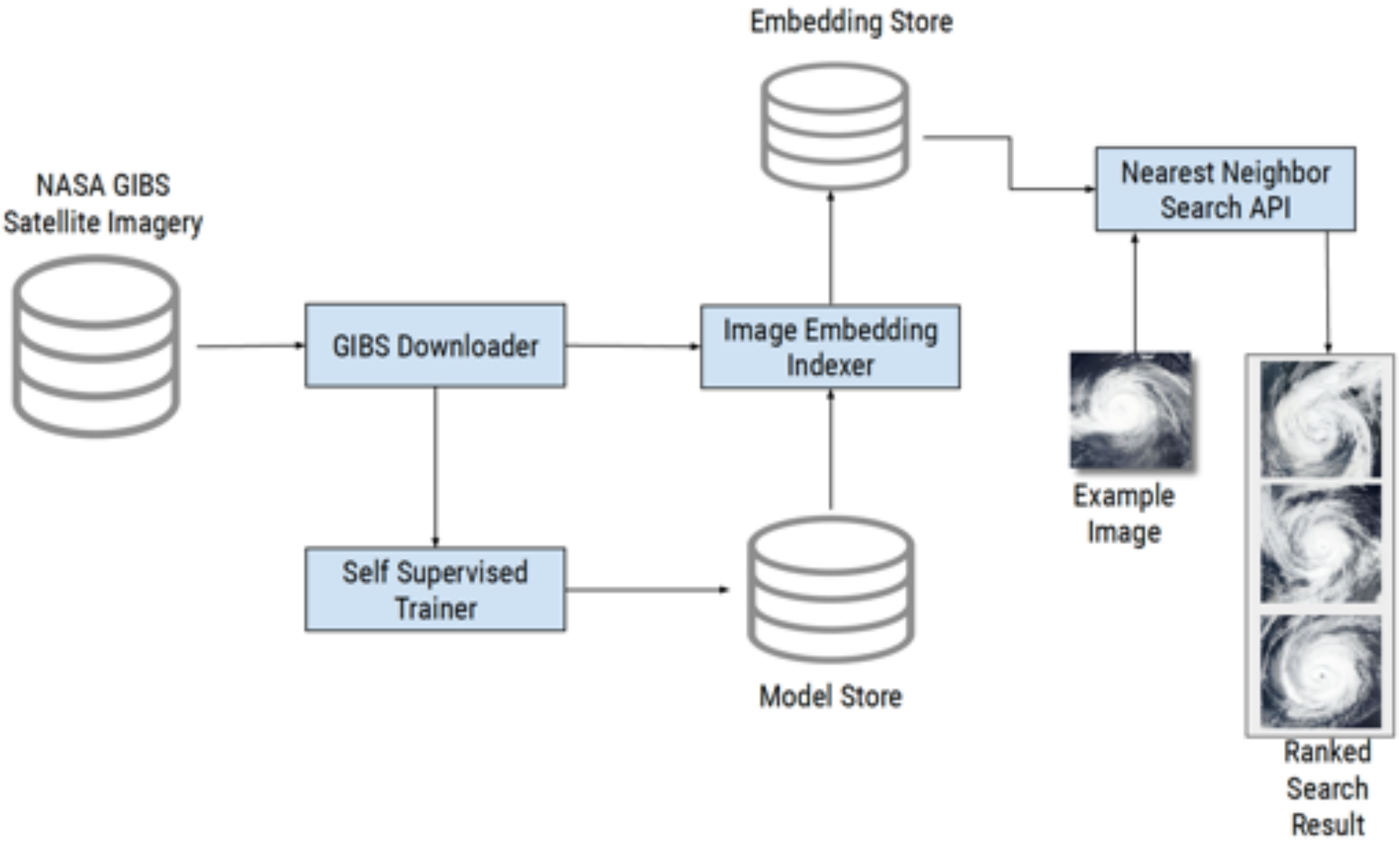
Phenomena Detection Portal

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

Types	Detections	Confidence score
03	98,627	89.61%
so far	and counting	on average

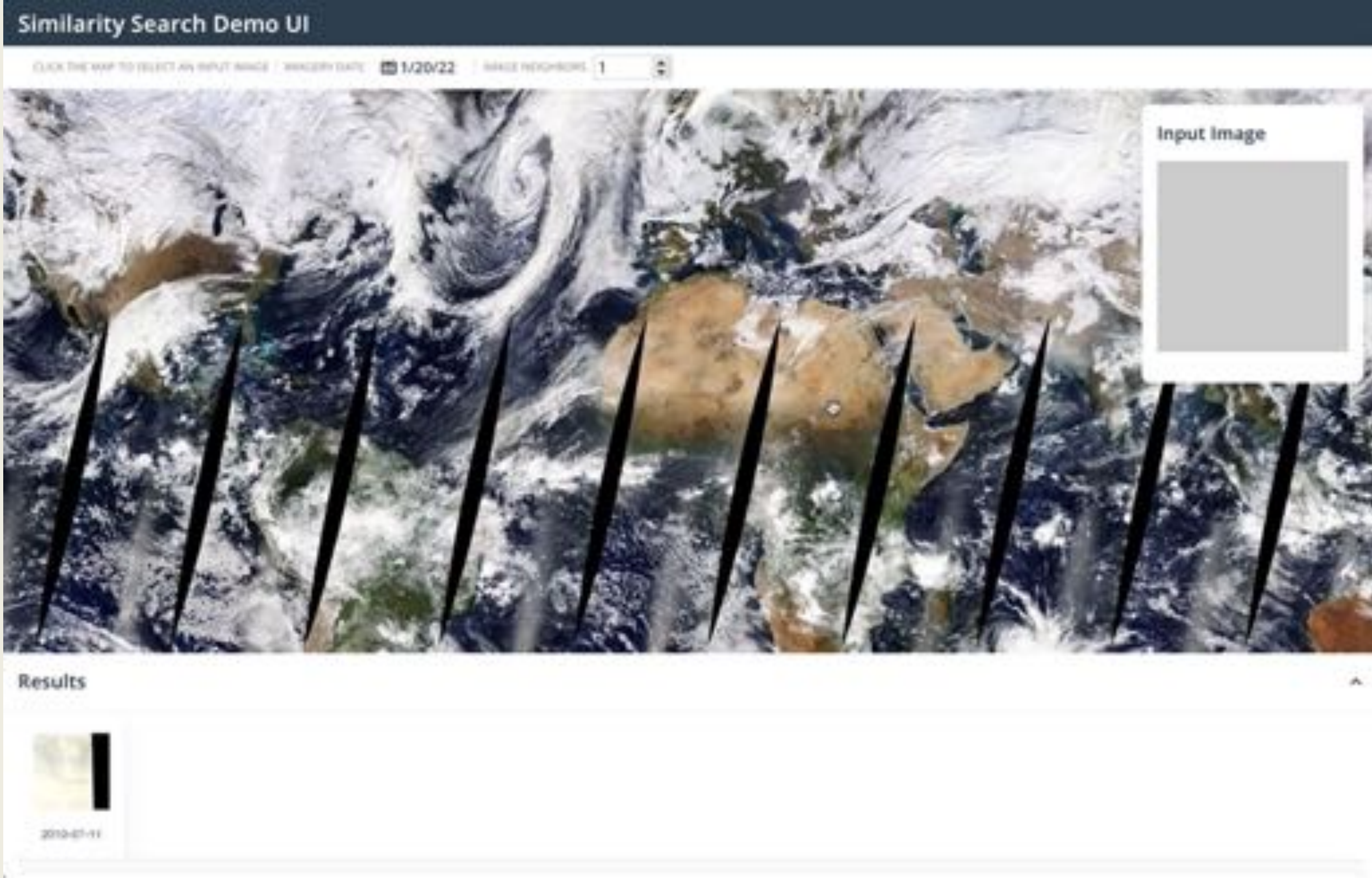
Start exploring | Learn more

Search by Example



Embedding Store
NASA GIBS
Satellite Imagery
Nearest Neighbor
Search API
GIBS Downloader
Image Embedding
Indexer
Example
Image
Self Supervised
Trainer
Model Store
Ranked
Search
Result

Scaled SSL



Augment data stewardship processes

Why?

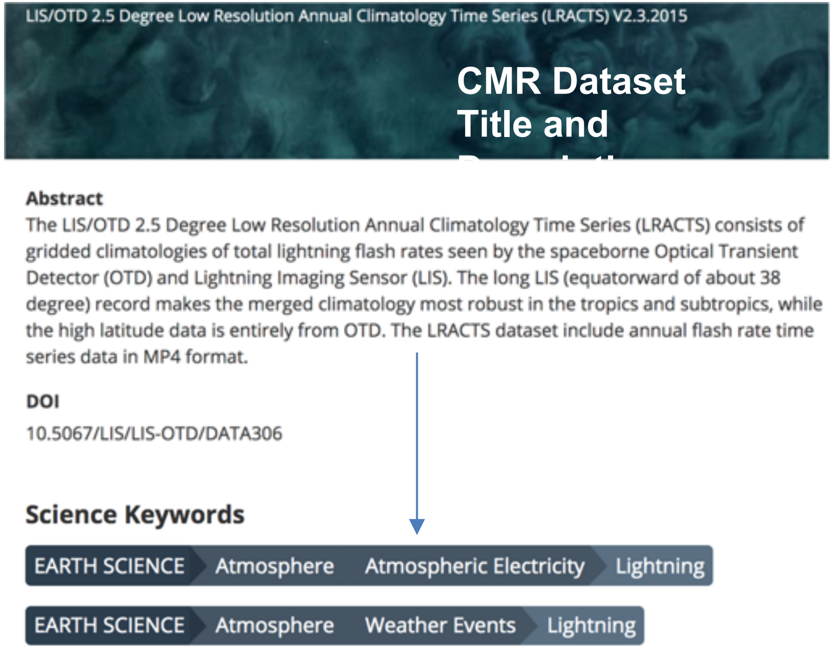
Assigning science keywords is currently a manual process, which is prone to human error and inconsistencies.

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



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	59	64	34	80	69	67	69	65	67	58
Earth and Space Science													1	24	28	25	42	88
Earth's Future												13	26	24	52	58	80	56
Eos, Transactions American Geophysical Union										46	37			1				
Geochemistry, Geophysics, Geosystems	88	207	266	173	195	143	228	174	164	203	192	247	196	154	185	567	64	29
GeoHealth																22	16	14
Geophysical Research Letters	1,155	1,436	1,518	1,688	1,700	1,533	1,559	1,290	1,005	1,234	1,058	1,154	1,200	1,388	1,491	1,390	1,507	1,136
Global Biogeochemical Cycles	132	136	124	120	75	66	94	66	83	64	96	76	72	76	75	74	23	21
Journal of Advances in Modeling Earth Systems								6	3	9	28	35	49	88	66	111	125	134
Journal of Geophysical Research							56				20							
Journal of Geophysical Research: Atmospheres	873	1,238	746	706	784	965	972	737	954	948	782	969	811	748	841	758	784	558
Journal of Geophysical Research: Biogeosciences				23	79	140	191	112	146	185	130	92	109	158	138	132	45	24
Journal of Geophysical Research: Earth Surface		52	47	95	84	145	130	113	134	132	141	117	92	92	93	59	44	30
Journal of Geophysical Research: Oceans	253	497	375	314	317	341	434	323	382	531	418	338	341	330	325	388	402	284
Journal of Geophysical Research: Planets	132	259	125	126	192	150	195	130	147	171	172	507	89	78	75	103	142	92
Journal of Geophysical Research: Solid Earth	345	683	365	319	377	435	436	376	508	450	378	297	316	307	314	388	126	96
Journal of Geophysical Research: Space Physics	439	561	498	525	475	447	589	533	756	854	694	490	503	541	592	542	466	446
Meteorological Applications								7	47	69	76	76	46	61	67	76	1	2
Paleoceanography	65	109	98	82	67	61	85	47	64	59	55	45	59	76	56	58		
Paleoceanography and Paleoclimatology																	24	26
Quarterly Journal of the Royal Meteorological Society									6	105		203	168	197	178			
Radio Science	108	132	146	114	94	122	91	108	59	136	79	52	69	70	109	93	94	51
Reviews of Geophysics	9	23	12	11	8	54					12	12	14	23	16	22	16	17
Space Weather		16	57	53	48	47	44	48	48	55	63	65	53	67	68	88	125	58
Tectonics	55	85	79	88	58	73	66	60	73	67	78	58	78	83	99	119	41	58
Water Resources Research	387	396	319	337	328	364	414	356	350	403	408	447	412	365	403	397	453	346

88,410

documents

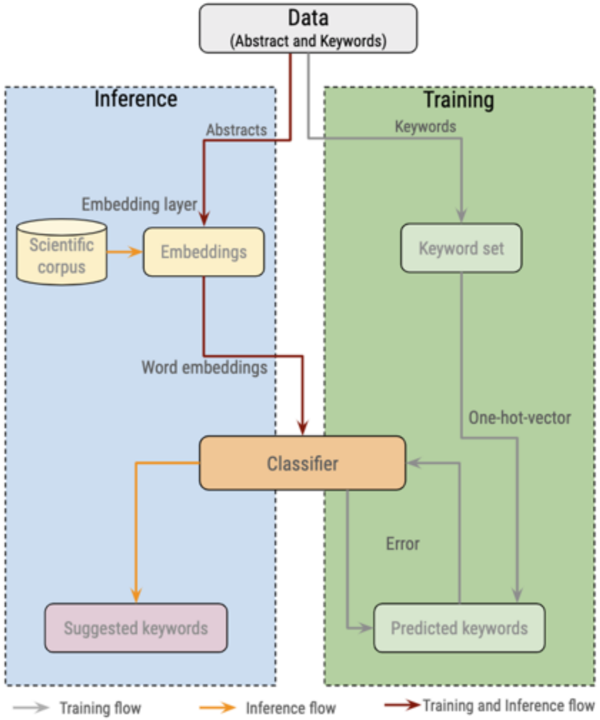
530 million

words

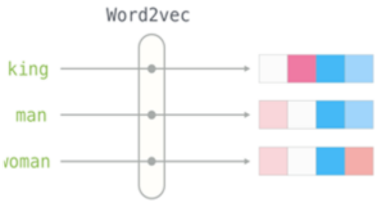
5.5 million

unique words

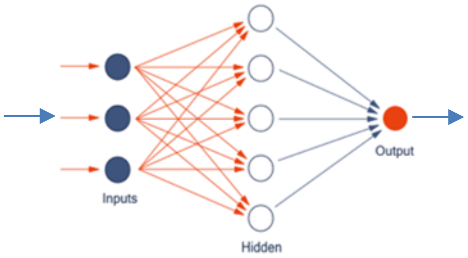
Automated keyword assignment



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



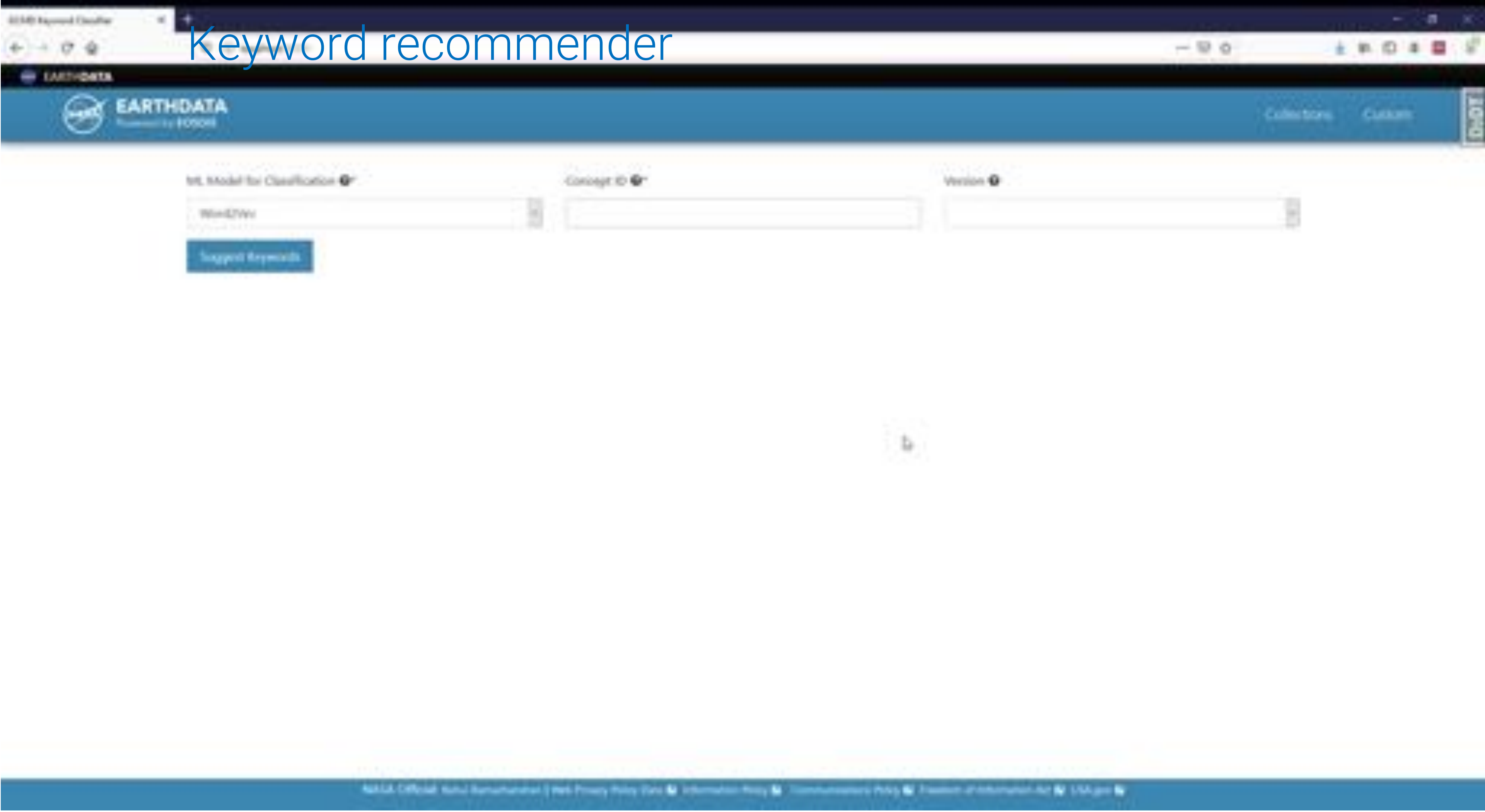
word2vec



classifier

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

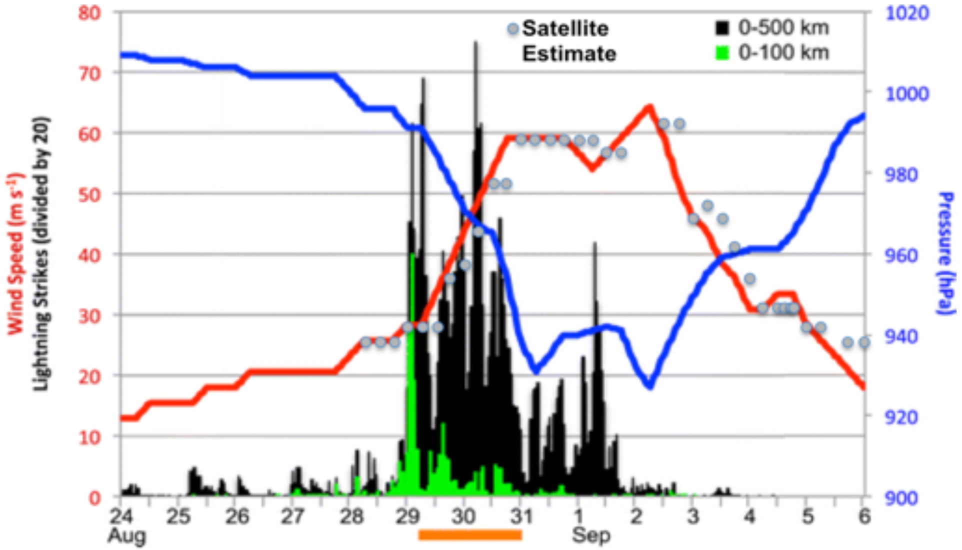
Keyword recommender



Hurricane intensity estimation system

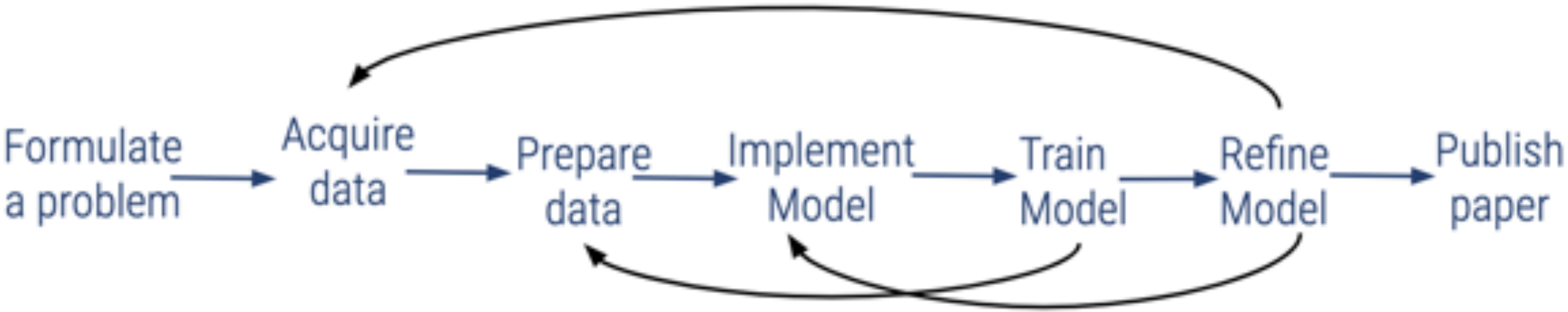
AI and satellite imagery to estimate hurricane wind speed

Hurricane Earl, 2010



Adapted from Stevenson et al. (2014). Time series of satellite-derived intensity estimates (circles) for Hurricane Earl (2010), added to best track intensities and lightning flash rate time series.

ML in literature



We have a model.....now what?

Going extra mile

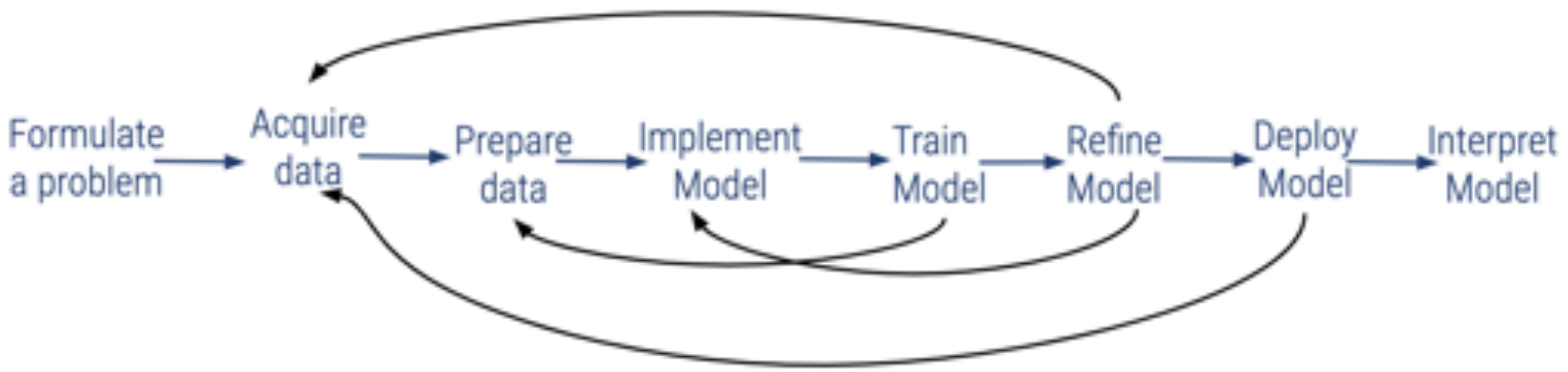
Interpretability + model inspection

Interpret prediction data – prediction output maybe just numbers

Questions:

Does the model confidence remain the same over time?
How do you maintain?
How do you complete the loop with new training data?

ML lifecycle - iterative



Deployment to production

Performance requirements

Metrics and baselines with initial models

Monitor over time

Back-testing

Model and software will change

Testing model changes on historical data

Run current production model to baseline performance

Run new models, competing for production

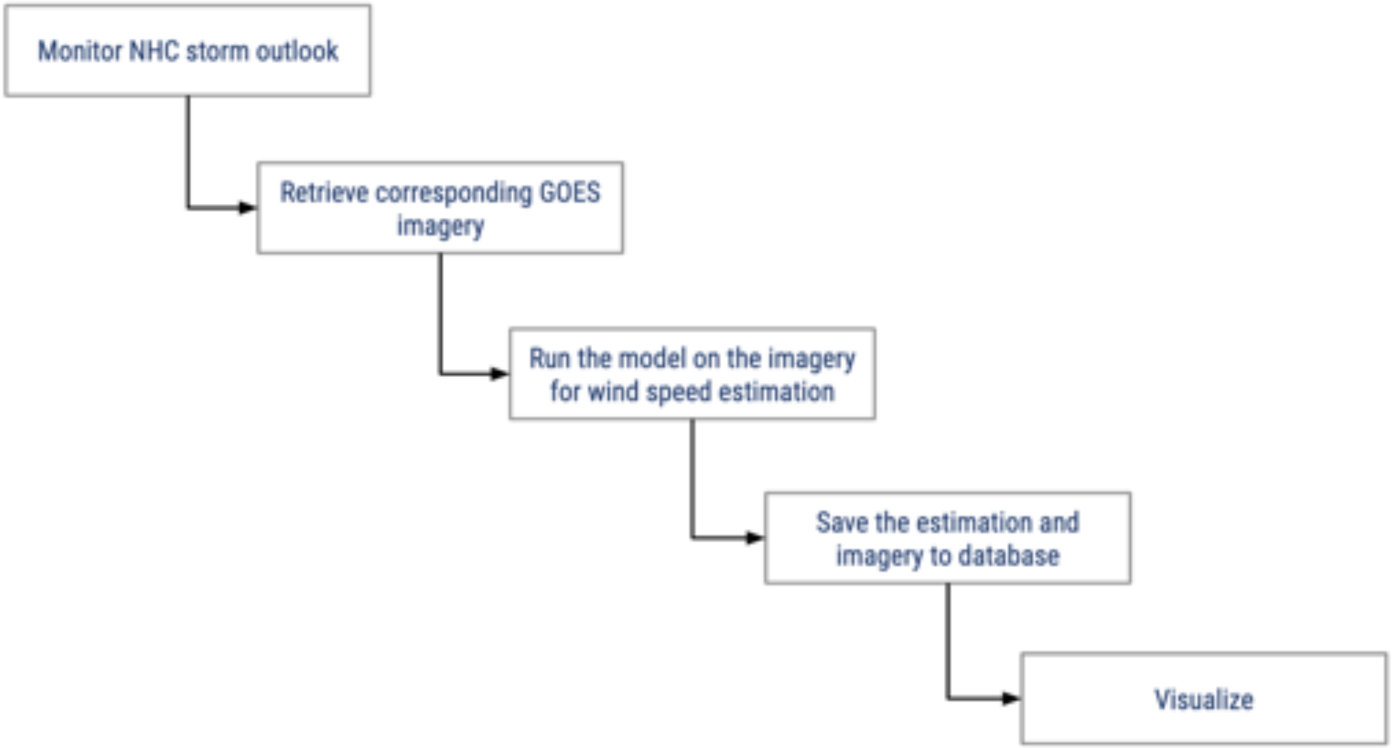
Now-testing

Testing of production model on latest data

Can we get early warning that the model may be faltering?

- Content drift: training data exploited by model are subtly changing with time

Workflow


Monitor NHC storm outlook
Retrieve corresponding GOES imagery
Run the model on the imagery for wind speed estimation
Save the estimation and imagery to database
Visualize

https://impact.earthdata.nasa.gov/hurricane/

Using community to advance model development

Data science competitions

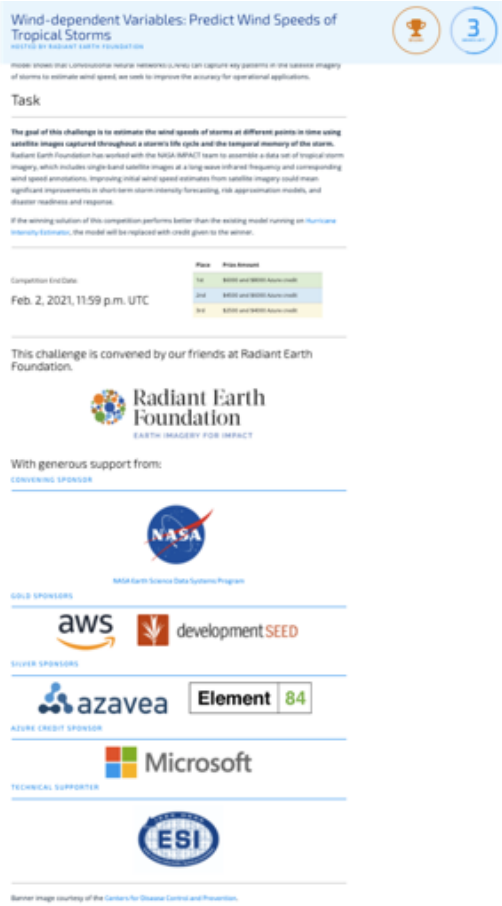
- Benchmark datasets and challenge problems have played an important role in driving progress in AI
  - Enables rigorous performance comparison
- Foster the learning of best practices
- Stimulate the abilities in problem-solving
- Encourage creativity and group work
- Give learners the chance to interact with new platforms and algorithms
- Citizen science

Data science competition

“Wind-dependent Variables: Predict Wind Speeds of Tropical Storms”

- Leverage community to enhance solution to existing problem using open data
- Test whether high-quality datasets produces better models via open competition

- **Industry** partnership
- **733** participants
- **2756** entries

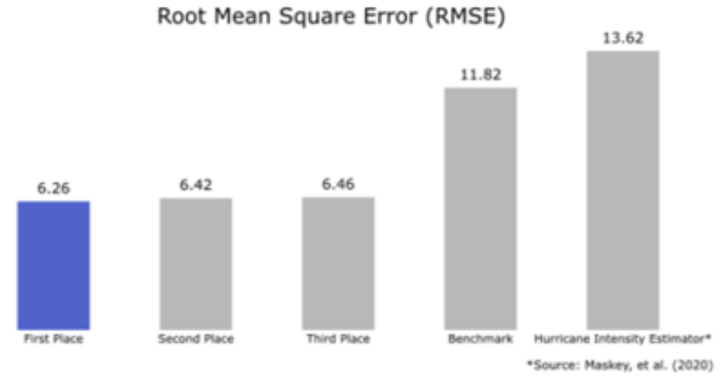


https://www.drivendata.org/competitions/72/predict-wind-speeds/

Data science competition

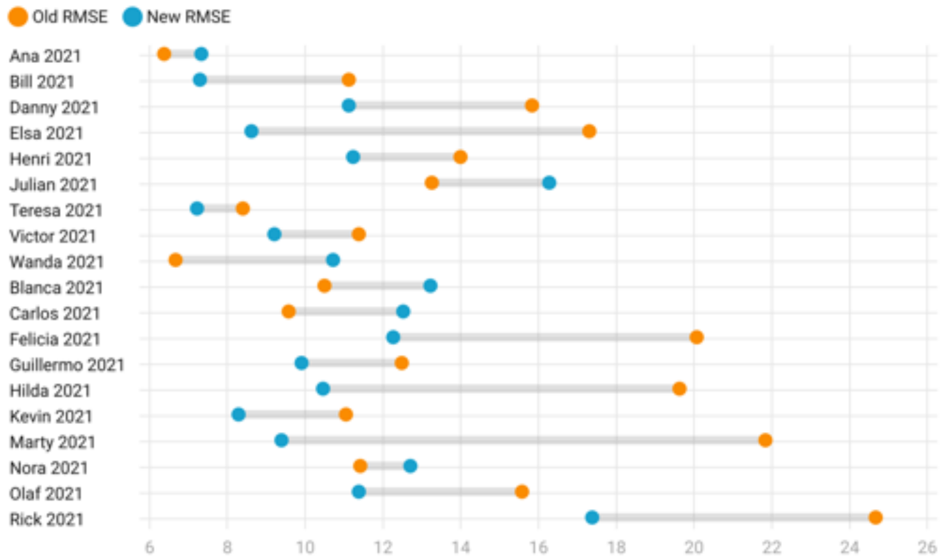
The Results

Over 700 participants stepped up to this important challenge, generating more than 2,700 entries. **Each of the top three models achieved at least a 50% reduction in Root Mean Square Error (lower is better) as compared to the existing model!**

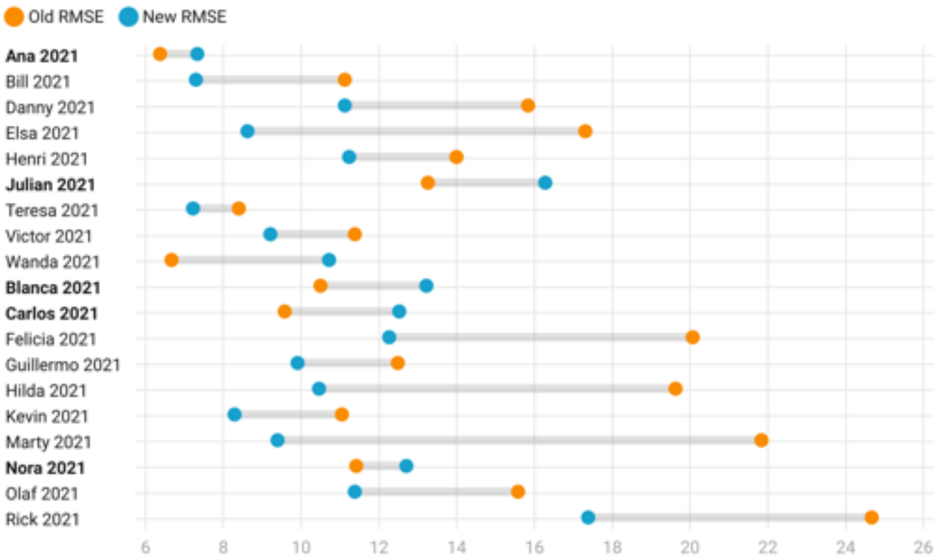


Winning solutions were able to take advantage of the relative timing of images in a storm sequence to produce targeted wind speed estimates based on temporal trends. As a result, these solutions can help to improve disaster readiness and response efforts around the world by equipping response teams with more accurate and timely wind speed measurements. All of the prize-winning solutions from this competition are linked below and made available for anyone to use and learn from.

New Model – in depth analysis



New Model – in depth analysis



Flood extent detection

Flood extent detection

- A major natural disaster
- Widespread damage – property, agriculture
- Displacement, insurance, long-term socio-economic consequences
- Causes:
  - Persistent rainfall
  - Severe storm
  - High-tides
  - Storm surge from cyclones

Problem

- Detecting flood extent is difficult
- Monitoring extent of flood events in-situ – hazardous to operate in a disaster zone
- Potential solution:
    - Remote sensing in conjunction with ML has been used in the community to monitor these events
- Need:
    - Large amounts of clean and labeled data

Data acquisition

Data collection

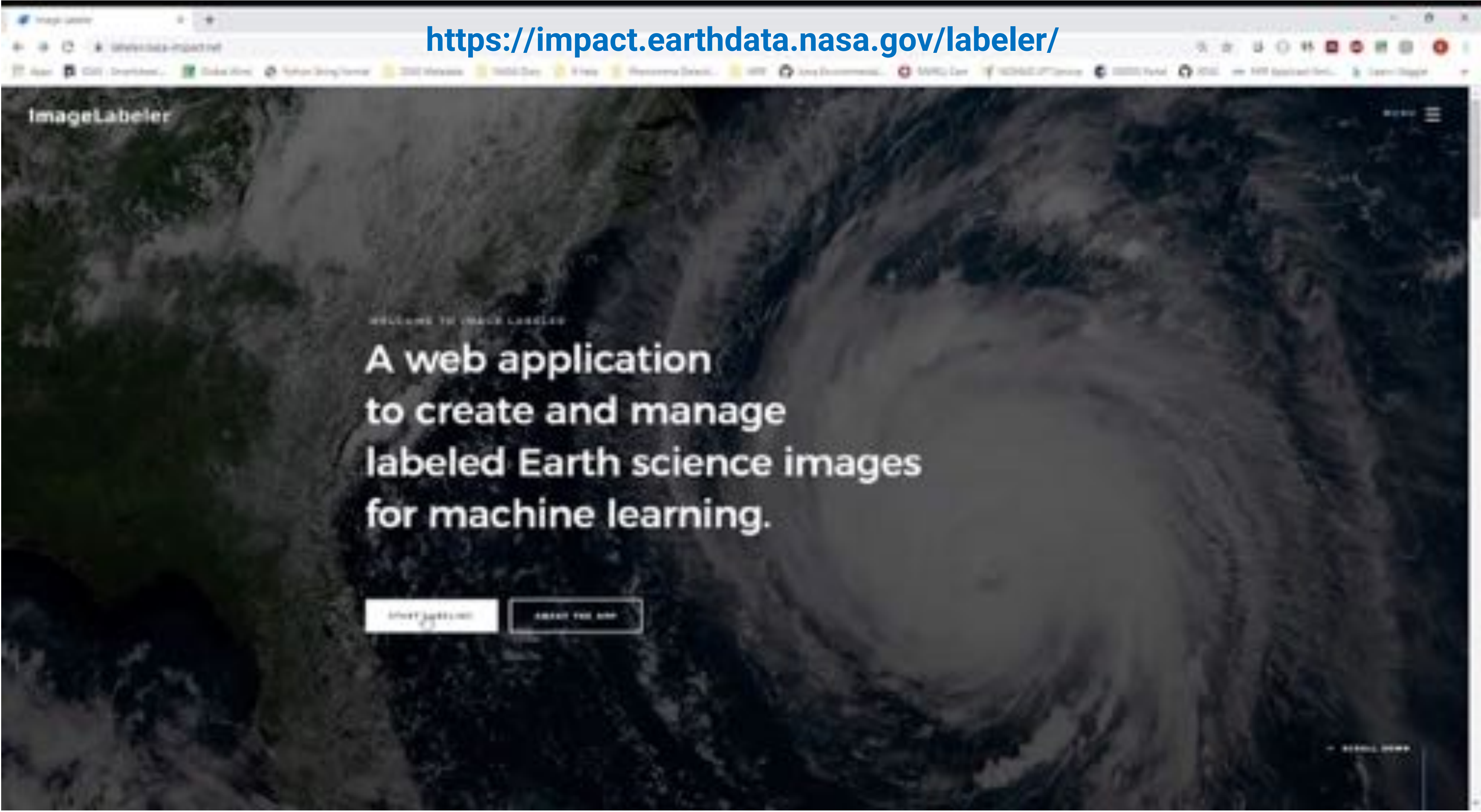
Data preprocessing

Data preparation

Data labeling

- 6 Atmospheric science/Earth science students
- 2 Domain scientists
- Training sessions
- Validation

https://impact.earthdata.nasa.gov/labeler/



Labeled data

~66k images (33k VV + 33k VH images including swath gap artifacts)

Native resolution : 5x20m

Train (24300):

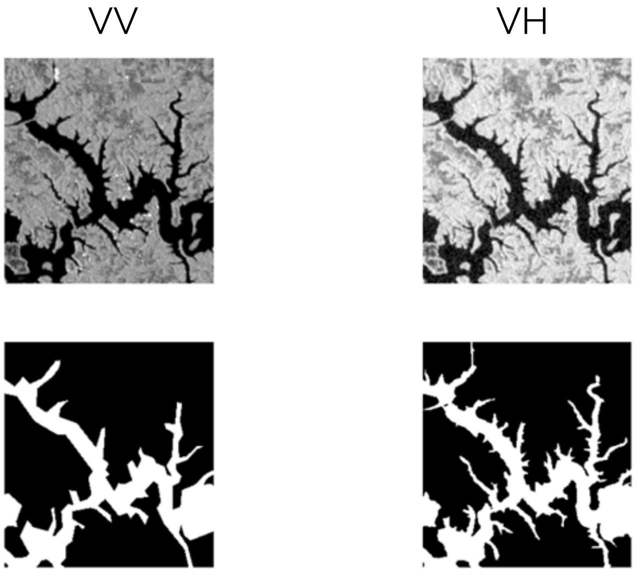
- Nebraska (1741 sq. km.) (~43 %)
- North Alabama (13789 sq. km.) (~43%)
- Bangladesh (7150 sq. km.) (~13%)

Validation (6500):

- Florence (7197 sq. km.)

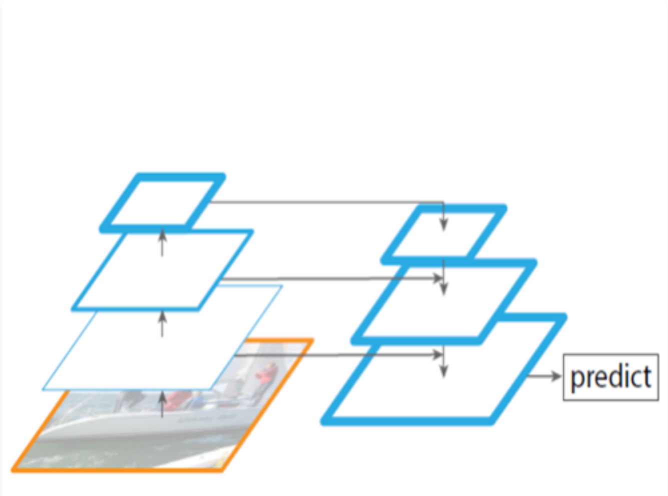
Test (1600):

- Red River North (6746 sq. km.)

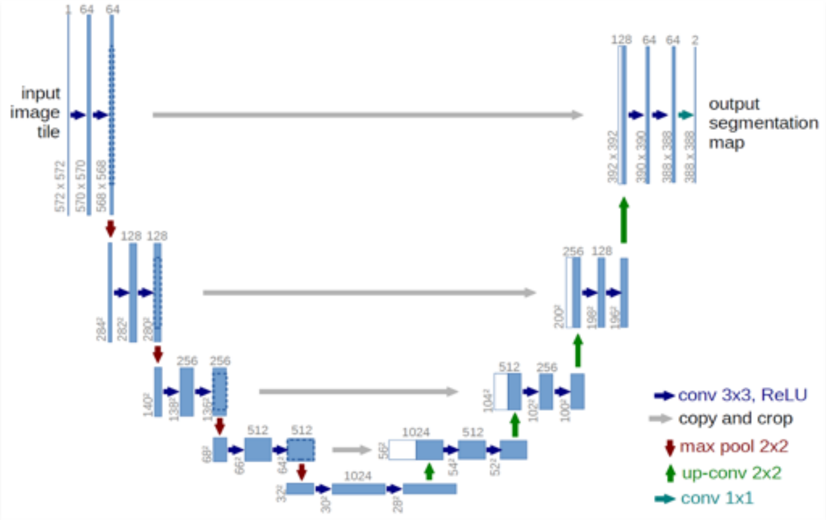


Benchmark model development

Baseline models: Vanilla FPN and U-Net (ResNet 50 encoder). random selection, 4:1 Train-Validation split



Feature Pyramid Network (Resnet 50 encoder)



U-Net (Resnet 50 encoder)

Benchmark model development

Visual Results

Sample segmentation

Leveraging citizen science for optimal model

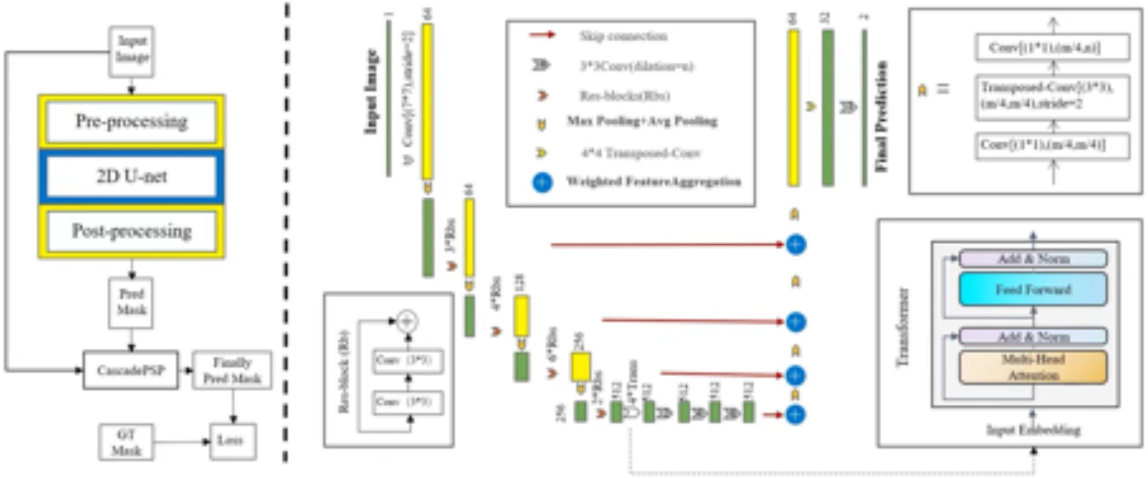
Finding optimal modal is an exhaustive task

ML Competition in collaboration with IEEE

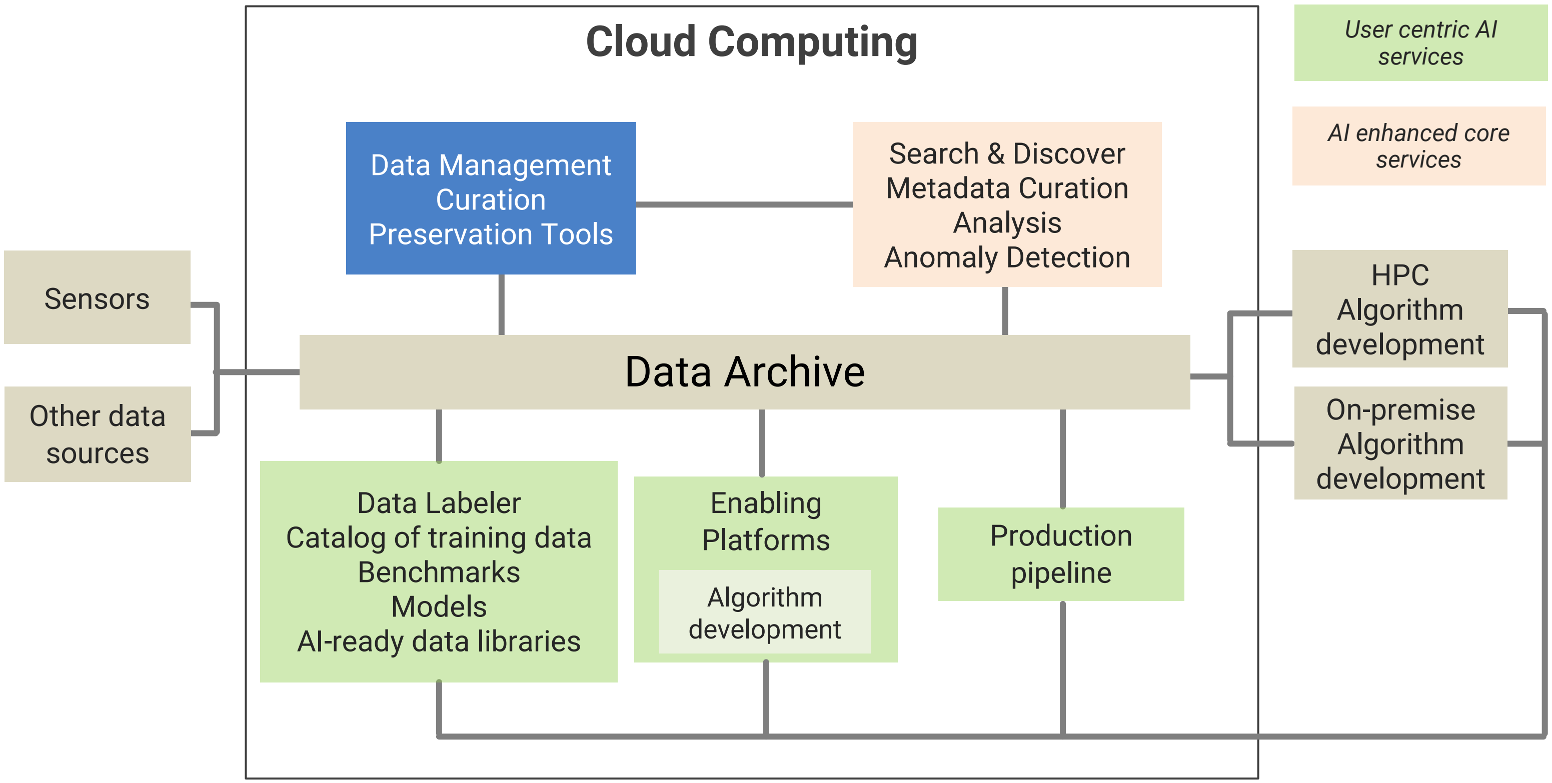
- 137 participants
- More than 200 submissions
- Codalab platform

- **Phase 1 (Development):** Participants are provided with training data (which includes reference data) and validation data (without reference data until phase 1 concludes) to train and validate their algorithms. Participants can submit prediction results for the validation set to the codalab competition website to get feedback on the performance from April 15 to May 14, 2021. The performance of the best submission from each account will be displayed on the leaderboard.
- **Phase 2 (Test):** Participants receive the validation set reference data for model tuning and test data set (without the corresponding reference data) to generate predictions and submit their binary classification maps in numpy array format from May 15 to June 30, 2021. After evaluation of the results, three winners will be announced on July 1, 2021.

Winning solutions



Team Arren, IOU: 0.7681

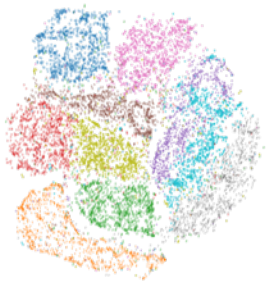

Cloud Computing
User centric AI services
AI enhanced core services
Data Management
Curation
Preservation Tools
Search & Discover
Metadata Curation
Analysis
Anomaly Detection
Sensors
Other data sources
Data Archive
HPC Algorithm development
On-premise Algorithm development
Data Labeler
Catalog of training data
Benchmarks
Models
AI-ready data libraries
Enabling Platforms
Algorithm development
Production pipeline

Summary

AI enhanced enterprise data systems

- AI approaches that can efficiently operate as a part of the core of large-scale systems
- Before AI can be widely used in critical enterprise data systems, we need new robust pipelines to systematically manage AI lifecycle
- A flexible architecture that allows software systems and AI algorithms to evolve to take advantage of emerging trends in hardware and software and rapid model deployment



Thank you.